

ToxiSpanSE: An Explainable Toxicity Detection in Code Review Comments

Jaydeb Sarker♣, Sayma Sultana♣, Steven R. Wilson◇, Amiangshu Bosu♣

♣ *Wayne State University, Detroit, Michigan, USA*

◇ *Oakland University, Rochester, Michigan, USA*

jaydebsarker@wayne.edu, sayma@wayne.edu, stevenwilson@oakland.edu, amiangshu.bosu@wayne.edu

| Toxicity (毒性)

- オンラインでの会話における否定的/反社会的なやりとり^[18]
e.g.) 攻撃的な言葉, ヘイトスピーチ, ネットいじめ
- 様々なオンラインコミュニティで問題になっている^{[2][3]}

| FOSS (Free and Open Source Software) コミュニティにおけるToxicity

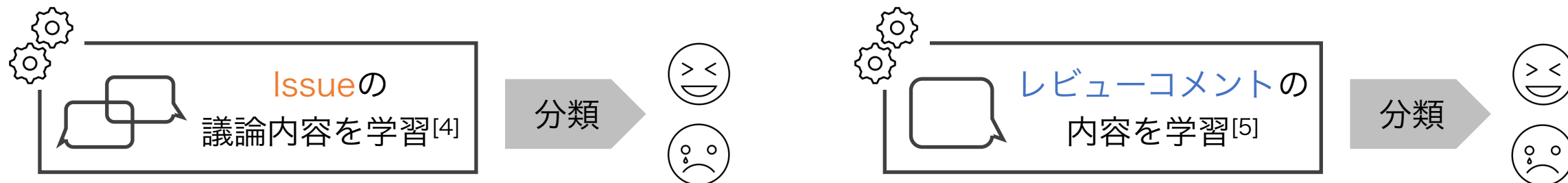
- ほとんどのFOSSコミュニティがToxicityを回避できない
 - 開発者のオンボーディングやメンタルヘルスへの悪影響を与える^{[4][28]}
- 市販のToxicity検出器はSEドメインに対応していないため信頼性が低い

[18] L. Aroyo, L. Dixon, N. Thain, O. Redfield, and R. Rosen, “Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions,” in *Companion proceedings of the 2019 world wide web conference*, 2019, pp. 1100–1105.

[2] C. Miller, S. Cohen, D. Klug, B. Vasilescu, and C. Kästner, ““did you miss my comment or what?” understanding toxicity in open source discussions,” in *2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)*. IEEE, 2022, pp. 710–722.

[3] A. A. Anderson, S. K. Yeo, D. Brossard, D. A. Scheufele, and M. A. Xenos, “Toxic talk: How online incivility can undermine perceptions of media,” *International Journal of Public Opinion Research*, vol. 30, no. 1, pp. 156–168, 2018.

SEドメインでのToxicity分類器[4][5]



Toxic Span（Toxicityの原因となるテキストの一部）の検出

- SemEval 2021でToxic Span検出のコンペが開催
→ BERTベースの検出手法が1位を獲得（F1-Score：70.83%）

本研究の目的

FOSSコミュニティ開発者間のToxic Spanを自動検出

- [4] N. Raman, M. Cao, Y. Tsvetkov, C. Kästner, and B. Vasilescu, "Stress and burnout in open source: Toward finding, understanding, and mitigating unhealthy interactions," in International Conference on Software Engineering, New Ideas and Emerging Results, ser. ICSE. ACM, 2020, p. TBD.
- [5] J. Sarker, A. K. Turzo, M. Dong, and A. Bosu, "Automated identification of toxic code reviews using toxicr," ACM Transactions on Software Engineering and Methodology, 2023.

データセット

データセット

- [5]で使用された19,651件のコードレビュー（CR）コメント
 - 3,757件のコードレビューコメントがToxicityにラベリング

データセットアノテーション

- Toxic Spanに対してアノテーションを付与
 - 男性1人，女性1人で実施

Character Span Offsets	CR Text
[10, 11, 12, 13, 14, 15]	Yeah that sucked , fixed done.
[39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 74, 75, 76, 77]	I think the formatting may have gotten screwed up (or Gerrit made it look ugly)
[]	below assignments also should be removed

データセット

アノテーター間の合意

- 付与したアノテーションが異なる場合, IAA (Inter-Annotator Agreement) を計算
- クリップエンドルフの α 係数で合意度を算出
 - 部分的な一致/不一致を考慮可能, $-1 \sim 1$ の範囲で1に近いほど一致度 (合意度) が高い

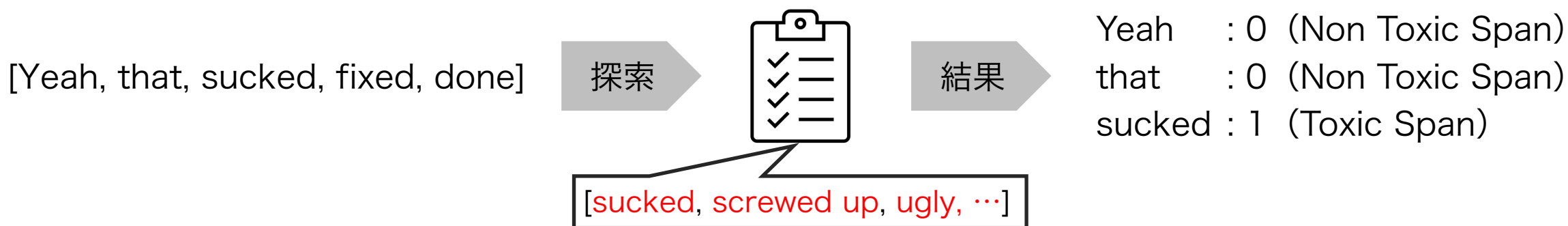
Rater	Text	Token Array	Character Spans
Rater1	if you think it sucks horribly , that's fine as long as we can fix it	[0,0,0,0,1,1,0,0,0,0,0,0,0,0]	[16-29]
Rater2	if you think it sucks horribly, that's fine as long as we can fix it	[0,0,0,0,1,0,0,0,0,0,0,0,0,0]	[16-20]
Final Label	if you think it sucks horribly , that's fine as long as we can fix it	[0,0,0,0,1,1,0,0,0,0,0,0,0,0]	[16-29]

- 3,757件全てでの α 係数は0.81 (ほぼ完全な一致)
- アノテーションが異なったものは, 話し合いで統一

アプローチ

辞書ベースモデル

- Toxic Spanに該当する語彙リストを作成し，各文の語彙をリストから探索
- 各トークンがToxic Spanかどうか出力



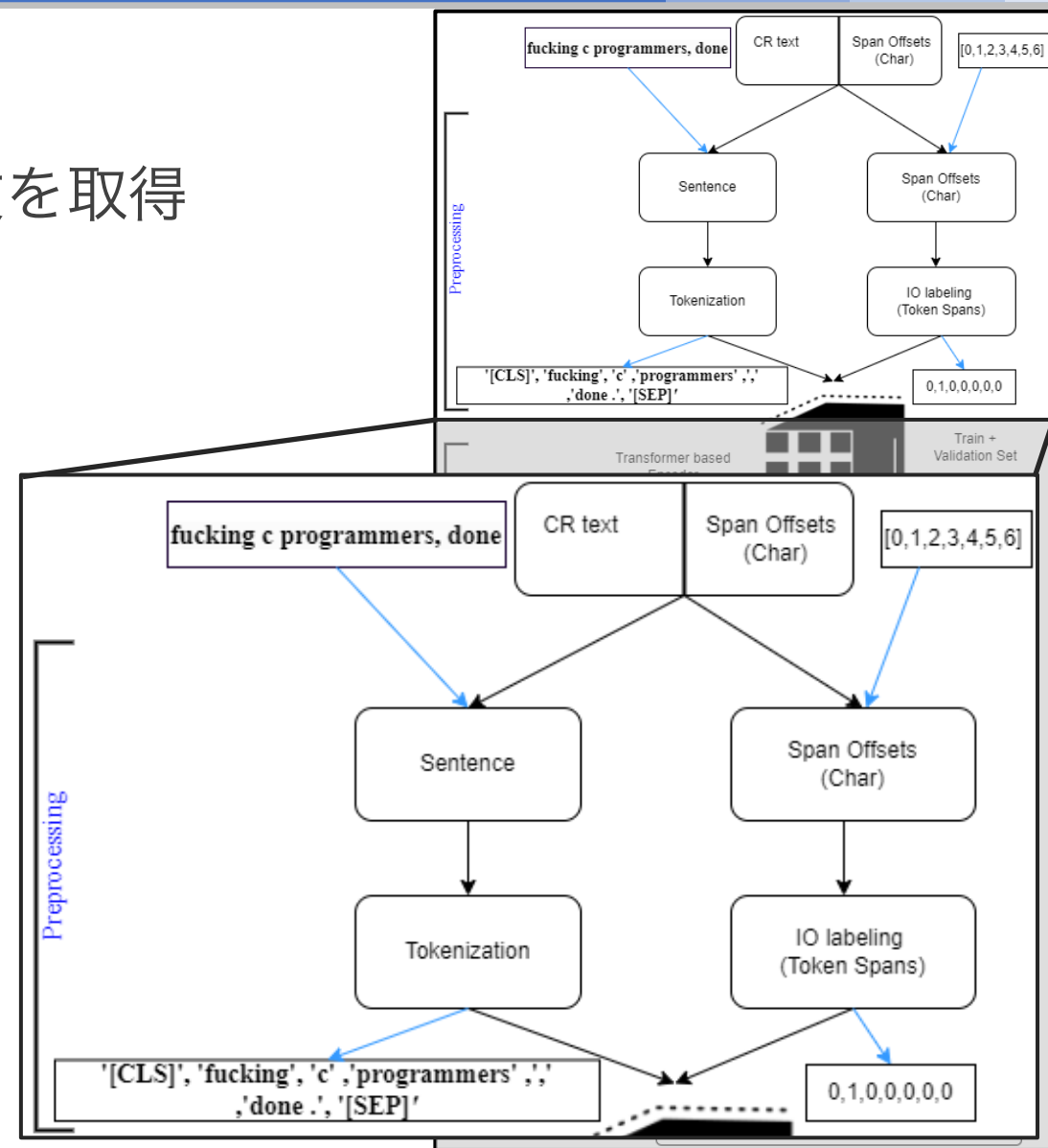
Transformerベースモデル

- BERT, DistilBERT, RoBERTa, ALBERT, XLNet を使用
 - 全てがエンコーダモデル

アプローチ

前処理

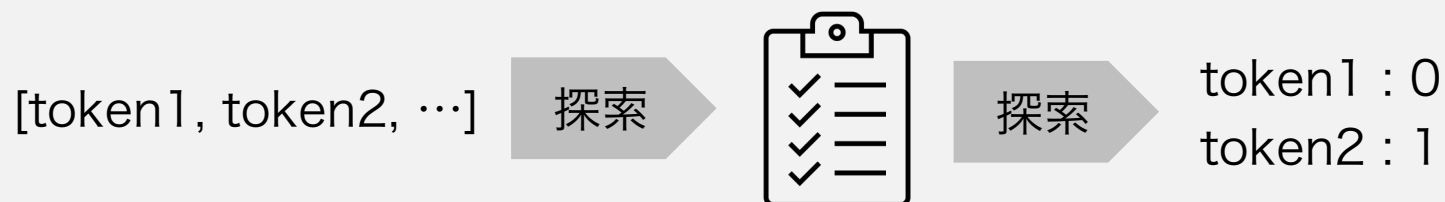
- 19,651件のCRテキストから39,438件の文を取得
 - 5,465の文がToxic Spanを持つ
- トークン化
 - 辞書ベース：NLTK
 - Transformerベース：Auto Tokenizer
- IOラベリング
 - 文のどの範囲にToxicityが含まれるかを識別



Sarkera^[5]の85個/Kurita^[66]の上位100個の冒涇的な言葉で辞書を作成



1. 辞書の作成



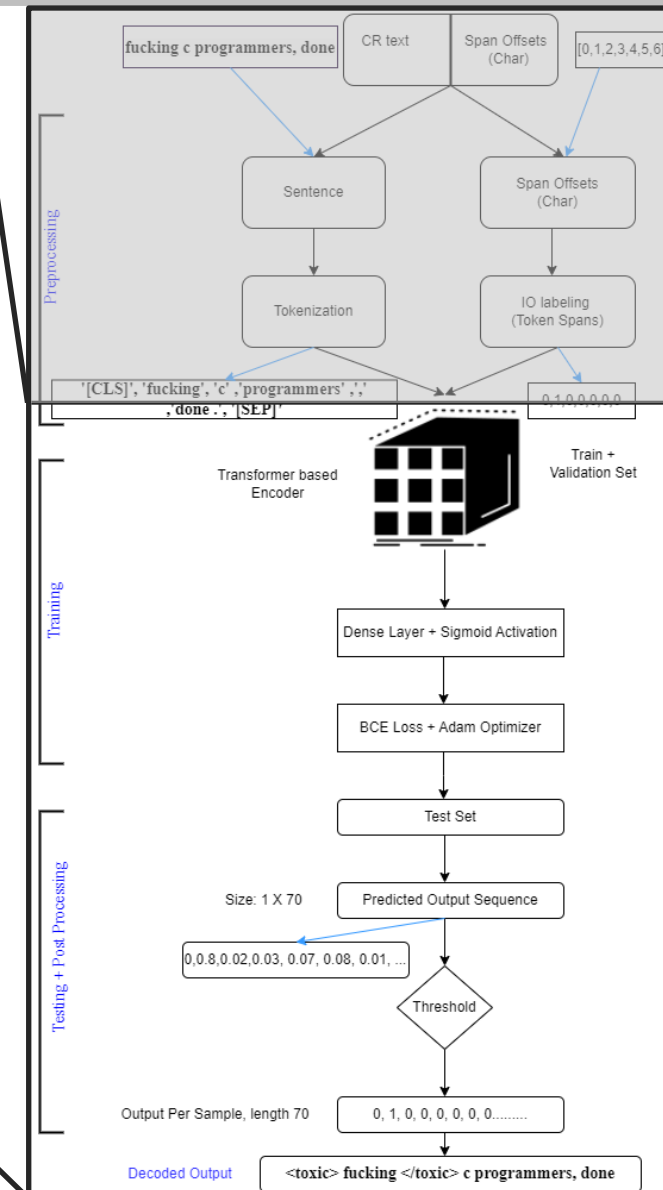
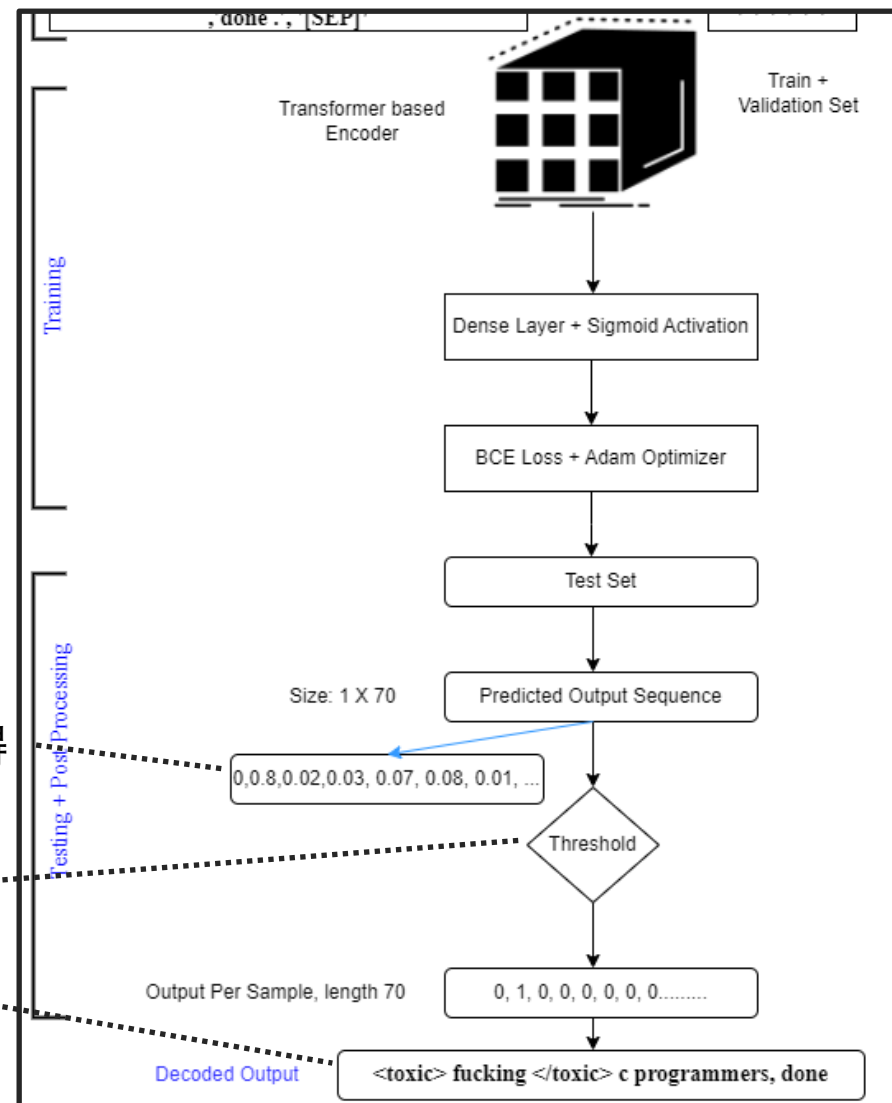
2. 辞書から各トークンがToxic Spanかそうでないか分類

[66] K. Kurita, A. Belova, and A. Anastasopoulos, "Towards robust toxic content classification," *arXiv preprint arXiv:1912.06872*, 2019.

アプローチ | Transformerベース

ファインチューニング

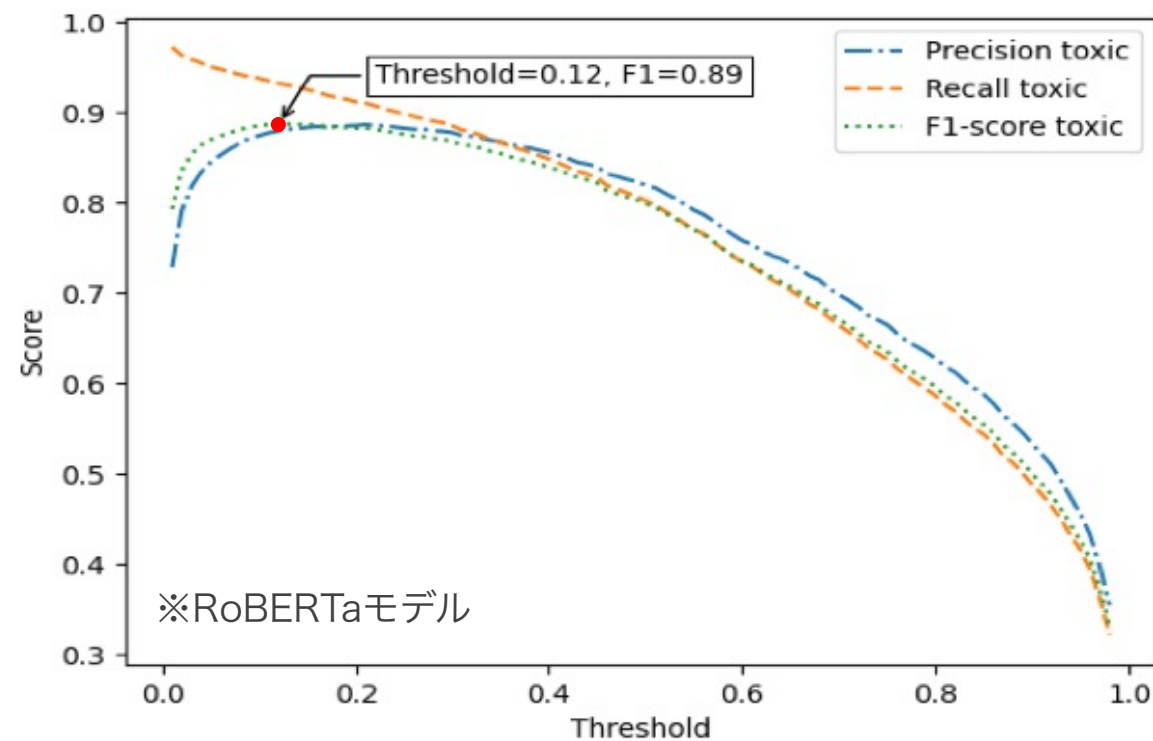
各トークンがToxic Spanである確率を取得
設定した閾値でToxic Spanかどうか分類
ユーザにはToxic Spanを強調表示で提供



実験

分類の閾値を0.01ずつ動かしながら最適な閾値を決定

- 10分割交差検証のそれぞれのテスト結果の平均
- F1値が最も高くなる値を最適な閾値として決定



予測結果

- データが不均衡なため、正/負例のどちらでも評価指標を計算

Models	Optimal Threshold	Non-toxic words			Toxic words		
		P_0	R_0	$F1_0$	P_1	R_1	$F1_1$
Lexicon-based	NA	0.95	0.95	0.95	0.75	0.67	0.69
BERT-base	0.15	0.95	0.95	0.95	0.87	0.89	0.86
RoBERTa	0.12	0.92	0.92	0.92	0.87	0.93	0.88
DistilBERT	0.17	0.94	0.94	0.94	0.85	0.89	0.85
ALBERT	0.11	0.92	0.92	0.92	0.85	0.89	0.85
XLNet	0.10	0.90	0.90	0.90	0.79	0.88	0.81

- RoBERTaが最も高い性能を示す
 - 他のTransformerベースのモデルも十分に高い性能

Toxic Spanの予測は辞書ベースよりもBERTベースモデルが優れている

誤分類の分析

<部分的な不一致>
完全には一致しないが、重複箇所がある

Error Types	GT Span	Predicted Span	Actual Text	Predicted Text
PD	[14,15,16]	[15,16]	rest seems like too generic name and it's not clear in code what the hell rest means.	rest seems like too generic name and it's not clear in code what the hell rest means.
PD	[1]	[1,2]	Damn grammar :-P	Damn grammar :-P
PD	[1,2]	[1,2,3,4,5]	O crap , hate me: do we still need this one?. PD	O crap, hate me : do we still need this one?.
FP	[]	[1]	FC related code should be removed.	FC related code should be removed.
FP	[]	[1]	stupid design on my part.	stupid design on my part.
FP	[]	[7]	As far as I understood, WTF::HashMap does'nt support it.	As far as I understood, WTF ::HashMap does'nt support it.
FN	[1]	[]	Evil spaces must die.	Evil spaces must die.
FN	[2,3,4]	[]	Your brain is deficient , please fix, also done.	Your brain is deficient, please fix, also done.

<偽陽性>

<偽陰性>

| Toxic Spanの選択は極めて主観的

- 928のサンプルでアノテーター間の意見の相違があった
- 同様のラベル付けタスクでは、多くの時間と明確な基準を設けるべき

| 辞書ベースアプローチは一般化できない

- 完全マッチングで判断するため見逃しが多い可能性

| Transformerベースモデルは信頼性が高く、説明可能性が高い

- Toxic Spanと予測したトークンを強調表示してユーザへ提供するため

| プラグインとして本手法を活用することで自省を促すことができる

他のタスクで既にやられてることをToxic Spanに当てはめただけ感が強い

- 感情分析系の論文でこういう手法の研究みたことある気がする

”Toxic Spanの選択は主観的“とか言ったら研究成り立たないのでは…

論文は読みやすくて自分が書く時の参考にしようとは思った