大規模言語モデルを用いた ソーシャルメディアにおける有害情報の検出と伝播経路の解明

廣中詩織（京都大） 土屋雅稔（豊橋技科大）吉田光男（筑波大）

背景と問い

背景：SNS 上の有害投稿が社会問題化
問い：1. 有害投稿は SNS 上で増加しているのか？
2. 増加している有害投稿の特徴は？その伝播経路は？

有害テキスト判定モデル

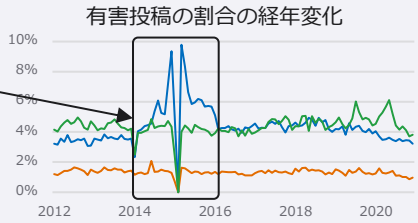
- LLM-jp Toxicity Dataset v2 を用いて学習
  - ✓ 有害ラベルは 1 or 0、件数は 3,847 件
  - ✓ 人間のアノテータによってラベルを付与
  - ✓ 有害ラベル：わいせつ、差別的、暴力的
- 事前学習済みモデル（cl-nagoya/ruri-v3-310m）を SetFit を用いてファインチューニング

分析対象データ

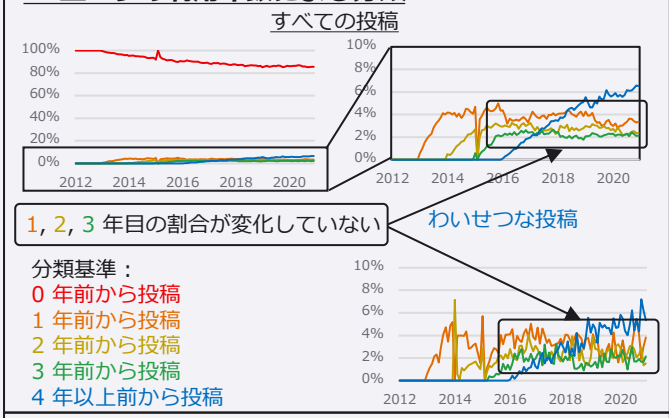
- 対象 SNS：X（旧 Twitter）
- 収集期間：2012年～2020年
- 合計件数：0.1% サンプリングして1,967,993 件
- 対象言語：日本語
- 収集方法：Twitter API v1.1

有害投稿の経年変化の分析

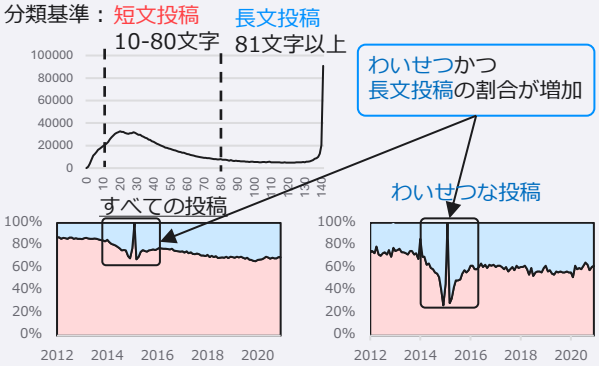
有害ラベル：わいせつ、差別的、暴力的
わいせつな投稿が2014-2015 年に増加
→ 増加の特徴を知るために 4つの方法で分類



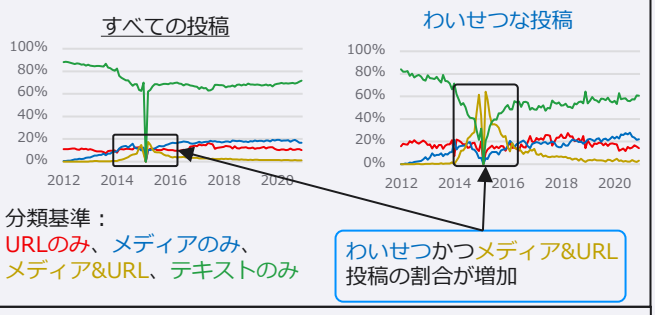
1. ユーザの利用年数による分類



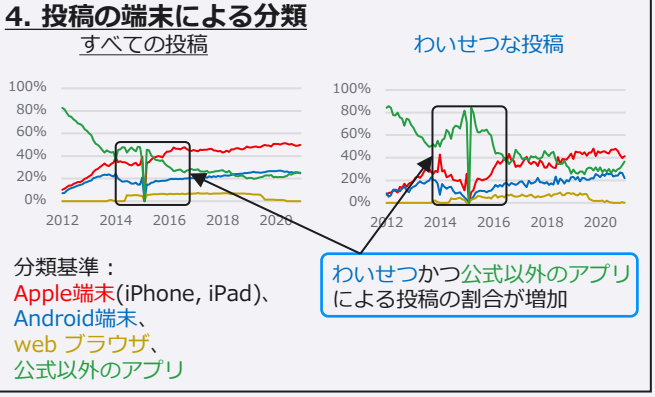
2. 投稿の文字数による分類



3. メディア・URLの有無による分類



4. 投稿の端末による分類



本年度の目標

- 有害テキスト判定モデルの分析・精度向上
  - 現在のモデルは，対象テキストを有害と判定した理由を提示できないため，有害テキストの特徴の分析が十分に行えない（その結果，伝播経路の分析も行えていない）
  - 現在のモデルは，学習データセットと分析対象データの性質（テキスト長，語彙，文体など）の乖離により，精度が不十分