令和 5 年度 特別研究報告書要旨

SNS における有害投稿の検出についての検討

Y200181 藤田 健真
指導教員:三好 力 教授

1.はじめに

現在、分散型 SNS に注目が集まっているが、分散型 SNS はその特性上、小規模団体での運営が行われている場合が多いため急激な環境の変化やユーザーからの多大な問い合わせに対応できない。そこで本研究では分散型 SNS のサーバー運営を助長することを目的として bert によるテキスト分類モデルを作成し、一部の有害と思われる投稿の検出を試みる。

2.提案手法

分散型 SNS の一種である Mastodon にて投稿を収集し、東北大学が公開している事前学習済み bert モデルに対してファインチューニングを行って研究で用いるモデルを作成する。“センシティブ発言”、“攻撃的発言”の 2 種類の投稿を検出の対象とし、センシティブ・攻撃的発言抽出モデル、センシティブ・攻撃的・その他の発言抽出モデル、センシティブ発言抽出モデル、攻撃的発言抽出モデル、の 4 つのモデルを作成する。作成したモデルは既存技術との性能比較、実際に Mastodon から目的の投稿が検出可能なのかの確認をテストデータを用いて行う。

3.実験 1

実験 1 では作成した学習モデルと既存技術を比較する(センシティブ・攻撃的・その他の発言抽出モデルについてのみ記載)。

表 1:作成モデルについてのパラメータ

精度	適合率	再現率	F 値
0.773	0.745	0.660	0.643

表 2:既存技術についてのパラメータ

精度	適合率	再現率	F 値
0.900	0.935	0.860	0.896

実験 1 にて入力したデータはセンシティブ発言 50 件、攻撃的発言 50 件、その他の発言 50 件で構成されている。作成モデルはセンシティブ発言 33 件、攻撃的発言 17 件、その他の発言 49 件を、既存技術はセンシティブ発言 47 件、攻撃的発言 43 件、その他の発言 50 件を正しく識別した。

4.実験 2

実験 2 では作成したモデルが Mastodon から目的とする投稿を抽出できるのか確認、性能評価する。

表 3:作成モデルについてのパラメータ

精度	適合率	再現率	F 値
0.850	0.512	0.600	0.529

実験 2 にて入力したデータはセンシティブ発言 17 件、攻撃的発言 18 件、その他の発言 223 件で構成されている。作成モデルはセンシティブ発言 11 件、攻撃的発言 5 件、その他の発言 183 件を正しく識別した。

5.考察

全体的な精度などの評価値については既存技術に劣る結果となったが、自作モデルが識別に成功していて、既存技術が識別に失敗している例もあり、自作モデルは利用目的によっては優れているといえる。

また実験 2 の結果より適合率や再現率の数値に不安が残るものの抽出対象としている投稿の検出に成功している。