2ZG-07

SNSにおける不適切投稿の検知

尾崎　航成　　向井　宏明　　松井　くにお
金沢工業大学　工学部　情報工学科

1. はじめに

幅広い世代へのインターネット普及に伴い，SNS(ソーシャルネットワークサービス)が広く利用されており，SNS上の人権侵害が増加した．2008年の人権侵害の件数は786件から，2018年は1910件へ増加した[1].

SNS上における人権侵害といっても様々な形があり，「名誉毀損」「侮辱」「脅迫」「ネットいじめ」「児童ポルノ」「ハラスメント」「差別」がある．どの分野での人権侵害でも共通していることは，Twitter， Facebook, Instagramなどの SNSなどのインターネットの便利な機能が誤用・悪用されて，人権侵害の道具として使用される[2]．SNS上の人権侵害が増加しており，これに伴いサイバー防犯ボランティアの負担も増加している．2011年のサイバー防犯ボランティア団体数は67団体から，2018年は244団体へ増加している[3].

日本国内における SNSの利用者は，年々増加しており，2020年末には SNS利用者数は7937万人，ネットユーザー全体に占める SNS利用率は78.7%に達する見通しである[4].

今後も SNSの利用者が増加し，SNS投稿数も増加が予想されるため，不適切投稿の監視自動化が求められる．現在は，2000 年代から第 3 次人工知能ブームが到来し，膨大な SNSのデータから不適切投稿を検知できるようになり，AI で SNS を監視するシステムに関する研究が行われている．本稿では主に SNS投稿データの中から人権侵害などに該当する不適切投稿を機械学習で検知する手法を検討し，評価を行ったので報告する．

2. 課題

SNS 上での人権侵害の増加に伴い，インターネット監視団体および監視者の負担が増加し，不適切投稿の人的監視が容易でなくなった．

従来の手法では，SNS 投稿を人手によりチェックし，不適切投稿の検知をして法務局人権擁護委員に投稿削除依頼を行っている．このため，投稿が削除されるまでの間に，膨大な時間を要する[5].

近年では，検索エンジンの発達により投稿されたデータに対してルールベースにより不適切投稿を検知する取組みが行われている．特定の禁止ワードをルールとしてルールベースを持ち，投稿データに対してパターンマッチングにより自動的に抽出される．その後，監視者によって投稿を削除するかどうかが判断される．その監視方法において，登録する禁止ワードは，時代とともに変化するため，常に更新が必要である．また，禁止ワード使用せずに他の言葉を使用しての人権侵害を行う投稿もある．この場合には，不適切投稿として扱われない．

Detection of inappropriate posts on SNS
Kosei Ozaki， Hiroaki Mukai， Kunio Matsui
Kanazawa Institute of Technology

3. 教師あり学習による不適切投稿の提案手法

SNS における不適切投稿の検知による提案手法を述べる．本研究では，不適切なニュアンスの投稿を学習させ，不適切投稿を検出できるように教師あり学習によるテキスト分類を用いた．

データベースでは，TwitterAPI を用いて Twitter から投稿を収集する．それらのテキストデータをデータベース(SQLite)に格納する．

構文解析では，CaboCha を使ってテキストの文法的な構造を解析する．SQLite に格納しているテキストデータに対して，CaboCha を実行して構文構造を解析し，出力結果を SQLiteに格納する．

検索エンジン(Apache Solr)は，全文検索システムである．Solr は，一般的な関係データベースよりも速くテキストを検索することができる． 本研究では，膨大なテキストデータ(SQLite)から学習の元となるデータを出力するために「文単位で検索する検索エンジン」を使用する．

教師あり学習によるテキスト分類は，Twitter の投稿文に対して，その投稿が，「不適切なニュアンスの投稿」か「それ以外か」の 2 つのカテゴリに分類する．教師あり学習は，あらかじめ手作業で学習データを作成し，その学習データからテキスト分類の規則を自動で生成する機械学習の方法を用いる．教師あり学習の概要を図 1 に示す．

教師あり学習には，学習フェーズと分類フェーズがある．学習フェーズでは，学習データから分類の規則性を自動で抽出し，学習モデルを作る．学習データには，それぞれのテキストに対して，SNS 監視者ならどう分類するかの見本が沢山格納されている．続いて分類フェーズでは，学習モデルを使ってテキスト分類を実行する．教師あり学習による分類フェーズの概要を図 2 に示す．

まず，テキストから BoW 特徴量を抽出し，ベクトルにする．BoW 特徴量は，ベクトルの各次元はある特定の単語を表し，ベクトルの値はその単語の出現回数を表す．これにより，どのような単語がテキストに含まれていたか表現する．次に，線形 SVM という機械学習手法を用いる． 線形 SVM は，このデータ群を直線で，カテゴリ「０」とカテゴリ「１」に分割する．この時それぞれのカテゴリの直線に最も近いデータからの距離が最大になるように直線を決定する．

この直線を決定するのは，「分類を行うにあたっての単語の重要度」を調整する．

自動で学習した学習モデルを用いることで，テキストを分類する．教師あり学習は，学習データを使用して正解のカテゴリを出力できるように単語の重要度を自動で調整するため，手作業で規則を記載することなく，テキスト分類を行うことができる[6].

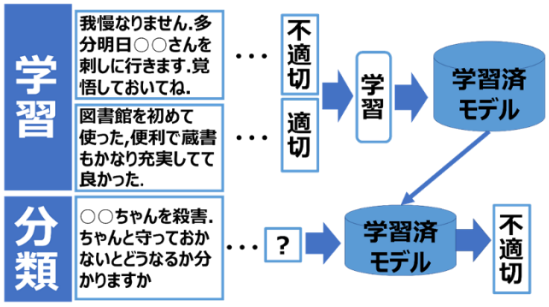

図１　教師あり学習の概要

図２　分類フェーズの概要

4. 検証

本研究では，TwitterAPI を用いて投稿を収集した．データ収集する際，人が嫌がるワードを検索ワードとして，「死ね」「消えろ」「嫌い」「罵り」「馬鹿」「クズ」「ゴミ」「ボケ」「アホ」「いじめ」のキーワードを含むそれぞれの文書に対して 1000tweet を収集した． 1 検索ワードにつき 1000tweet として，10 検索分の 10000tweet を収集した．

教師あり学習によるテキスト分類を実装するにあたって学習データが必要である． 学習データ作成において，不適切なニュアンスの投稿をカテゴリ「１」とラベル付け手作業でする．投稿文から「名誉毀損」「侮辱」「脅迫」「ネットいじめ」「児童ポルノ」「ハラスメント」「差別」といった人権侵害をする SNS 投稿を不適切なニュアンスの投稿としてカテゴリ「１」としてラベル付けする．それ以外を適切投稿としてカテゴリ「０」とする．一文一文目視で確認し，出力されたデータのうち正解である文のカテゴリを「０」から「１」に変更すると学習データが作成できる．

教師あり学習によるテキスト分類における評価方法について述べる．ラベル付を行った，学習データに対して，冒頭の 8000 件の tweet を学習用データとして残りの 2000 件の tweet を検証用データとする．

教師あり学習によるテキスト分類の分類結果において，「検証用データの正解ラベル」と「機械学習の分類ラベル」を比較し合否判定を表示する．合否判定の表示には，4 パターンある．表２の「分類結果 （ラベル付）」において，「「適切投稿」に対して「適切投稿」だと判断」を[0,0]，「「不適切投稿」に対して「不適切投稿」だと判断」を[1,1]，「「適切投稿」に対して「不適切投稿」だと判断」を[0,1]，「「不適切投稿」に対して「適切投稿」だと判断」を[1,1]とする．教師あり学習によるテキスト分類の性能を見極めるため正解率を算出した．算出式は（１）の様に表す．

$$正解率=\frac{[0,0]+[1,1]}{[0,0]+[1,1]+[0,1]+[1,0]} \qquad (1)$$

本研究では，教師あり学習によるテキスト分類の性能評価をするために，交差検証を行った．本研究における交差検証の手法は，K-分割交差検証を行った．

K-分割交差検証において冒頭の 8000 件の tweet を学習用データとして残りの 2000 件の tweet を検証用データとして，学習用データと検証用データの入れ替えを 5 回行い，5 回交差検証を行った．

各交差検証における学習用データ，検証用データに「適切投稿」，「不適切投稿」があるか表１に示す．

表１　学習データ

	学習用データ		検証用データ	
検証	適切投稿	不適切投稿	適切投稿	不適切投稿
1	4944	3056	705	1295
2	4639	3361	1010	990
3	4409	3591	1240	760
4	4293	3707	1356	644
5	4311	3689	1338	662

教師あり学習によるテキスト分類における交差検証の検証結果を表２に示す．正解数の平均値は，1940.2 であり，正解率の平均値は，0.9701 である．交差検証の結果として，正解数，正解率の観点から特に大きなばらつきはなく，機能評価は，良いといえる．

表２　交差検証結果

	分類結果 (ラベル付)				評価	
検証	[0,0]	[1,1]	[0,1]	[1,0]	正解数	正解率
1	1314	643	24	19	1957	0.9785
2	1340	622	16	22	1968	0.9840
3	1225	735	15	25	1960	0.9800
4	982	966	28	24	1848	0.9740
5	610	1258	95	37	1868	0.9340

5. おわりに

本研究では，機械学習を用いて不適切投稿を検知する手法を検討し評価した． 人権侵害をする投稿にて頻繁に使用される単語を含むものを不適切投稿としてラベル付し学習モデルを作成し教師あり学習による不適切投稿の検知を評価した．機械学習による平均正解率は 0.97 と高い精度であった．

参考文献

[1] 平成 30 年における「人権侵犯事件」の状況について
http://www.moj.go.jp/content/001288006.pdf
2019 年 12 月 21 日参照
[2] 平成 30 年 12 月末におけるサイバー防犯ボランティア団体数等
https://www.npa.go.jp/cyber/pdf/h30_volunteer.pdf
2019 年 12 月 21 日参照
[3] 佐藤佳弘「インターネットと人権侵害」,武蔵野大学出版会,2016 年,p.18
[4] ICT 総研「2018 年度 SNS 利用動向に関する調査」
https://ictr.co.jp/report/20181218.html
2019 年 12 月 21 日参照
[5]法務省「インターネットを悪用した人権侵害をなくしましょう」
http://www.moj.go.jp/JINKEN/jinken88.html
2019 年 12 月 21 日参照
[6]柳井 孝介, 庄司 美沙,「Python で動かして学ぶ 自然言語処理入門」, 株式会社 翔泳社, pp.189-199, 2019.