6W-03

fastText を用いた電子掲示板の高リスクな投稿/カテゴリの検知と診断

栗原　優太†　　斉藤　和巳†

†神奈川大学大学院 理学研究科

1　はじめに

不特定多数が利用する電子掲示板のコンテンツモデレーションでは，争いや法的問題の原因となる投稿を早急に把握することが重要である．

本稿では，日本国内で人気が高いスレッドフロート型の電子掲示板を対象に，価値観や思想の衝突から言い争いになりやすいと言われている「政治や宗教等」，名誉毀損や誹謗中傷のような法的問題に発展することが多い「炎上やアンチ等」の要素を含む，掲示板の運営を行っていく上で高リスクな投稿/カテゴリを，fastText[1]を用いた文章分類器によって，自動で検知・リスク指標の算出による診断を行う手法を提案する．検知の精度の評価と，診断の妥当性を確認する実験により有効性を検証する．

2　関連研究

本稿に一部含まれる「炎上の検知」に関して，SNS(Twitter)は大西ら[2]が，動画コメント(ニコニコ動画)は竹内ら[3]が，ライブ配信(Youtube Live)は齋藤[4]らが，それぞれ手法を提案している．しかしながら，電子掲示板を対象とした研究，炎上に加えて，政治や宗教等にも着目した研究は，筆者らが知り限りでは，行われていない．また，投稿やカテゴリを対象にリスク指標を算出する試みも行われていない．

3　電子掲示板のデータ構造・対象カテゴリ

3.1　対象とする電子掲示板のデータ構造

スレッドフロート型の電子掲示板では，何らかのテーマやジャンルを持つ掲示板(板)の中に，個別の話題を扱うスレッド(記事)が多数投稿される．そのスレッドの中に，レス(コメント)が最大 1000 件まで書き込まれるという構造を持つ．本稿では，記事(スレッド)の投稿を最小単位，複数の掲示板をまとめたカテゴリを最大単位とする．

3.2　検知と診断の対象とするカテゴリ

日本国内の電子掲示板で人気があるエンターテイメントや趣味に関係するジャンルの中から，「政治・宗教等」と「炎上・アンチ等」に当てはまらないものを幅広く選定し，5 種類の中間カテゴリに分類した．これらのジャンルを検知・診断の「対象カテゴリ」とする．各カテゴリの詳細を表 1 に示す．

表 1:各カテゴリの詳細

カテゴリ	詳細
政治や 宗教等	ニュース,国際情勢,金融,経済,疫病, 選挙,政策,法律,国防,政治思想,テロ, 特定の宗教,特定の政党,特定の国家, ジェンダー,民族,差別,犯罪,警察等
炎上や アンチ等	特定の人物(芸能人,動画配信者, SNS ユーザー,スポーツ選手,作家, ブロガー,その他一般人等)に対する 炎上,アンチ行為,監視活動,攻撃等
対象とする カテゴリ郡	以下の 5 種の中間カテゴリ ・二次元（アニメ,漫画,ゲーム等） ・グルメ(料理,お菓子,ドリンク等) ・スポーツ(野球,サッカー,テニス等) ・電化製品（家電,スマホ,AV 機器等） ・その他趣味(手芸,園芸,時計,煙草, ペット,模型,野鳥観察,写真撮影, お絵描き,自動車,自転車,おもちゃ等)

4　提案手法

4.1　投稿データの加工と文章分類器の作成

「政治や宗教等」,「炎上やアンチ等」,「対象カテゴリ」の各カテゴリに属するジャンルを扱う掲示板から，文章分類器の学習に用いるスレッドを収集する．全スレッドのレスに対して，記号や数字,レスアンカーの除去等の下処理を行う．レスの文章中の URL について，高リスクなカテゴリで頻繁に話題となるウェブサイトが存在し，識別に有効であるため,ホスト名とドメインから記号を取り除き，一つの英単語として扱う．下処理を終えたら，Mecab と ipadic-NEologd を用いて，形態素解析と分かち書きを行う．スレッド内のレスを全て結合して一つの文章集合に加工，それにカテゴリのラベルを付与して，教師あり学習に使用できる加工スレッドデータを作成する．

加工したスレッドデータと fastText を用いて，教師あり学習を行い，文章分類器を作成する．この際，個別スレッドが各カテゴリである確率を 0～100%の範囲で算出するために，ハイパーパラメータの損失関数を ova に指定する．

4.2　文章分類器を用いた検知と診断の方法

実際に,検知・診断を行うカテゴリの掲示板について，4.1 節と同様にスレッドを収集し加工する．

高リスクな個別スレッドの検知は，文章分類器を用いて,カテゴリを推定することで行う．

個別スレッド/カテゴリに対するリスク指標の算出は,文章分類器が算出する「個別スレッドが各カテゴリである確率(0～100%)」を用いる．個別スレ

Detection and diagnosis of high-risk post/category on electronic bulletin boards using fastText
†Yuta KURIHARA　†Kazumi SAITO
†Kanagawa University

ッドiのカテゴリが「政治・宗教等」である確率をp_i，「炎上・アンチ等」である確率をa_i，判定したいカテゴリ全体のスレッド総数をnとする．個別スレッドのリスク指標は$h_i = p_i + a_i$，特定のカテゴリの「政治・宗教等」指標，「炎上・アンチ等」指標，リスク指標のそれぞれは

$p = \frac{1}{n}\sum_{i=1}^{n} p_i$，$a = \frac{1}{n}\sum_{i=1}^{n} a_i$，$h = p + a$

として算出する．値の範囲はp_i, pが0～100，a_i, aが0～100，h_i, hが0～200となる．これらの指標を参照し，個別スレッドとカテゴリ全体に，高リスクな要素がどの程度含まれるのか診断をする．また，これらの指標に対して，任意の閾値を設定し，高リスクなカテゴリの検知を行う．

5　実験と結果

5.1　学習に用いる訓練/テストデータの準備

複数のスレッドフロート型の電子掲示板郡サービスから，3.2節で定義したカテゴリに該当する掲示板のスレッドを，投稿日時2004年1月～2024年1月の範囲で,18104件収集した．

収集したスレッドの中で，話題が掲示板のカテゴリから著しく逸脱している・書き込み数が数件で成立していない・スパムに該当するものは，目視で確認して除外した．また，ホールドアウト法によって,精度評価を行うために，8:2の割合で，訓練/テストデータに分割した．

5.2　評価に使用する文書分類器の作成

訓練データとfastTextを用いて，文書分類器を作成した．学習時のハイパーパラメータについて,学習率は0.5，エポック数は500，単語のnグラムは3，損失関数はova，単語ベクトルの次元数は300,バケット数は200000，他はデフォルトに指定した．学習は12時間で完了した．

5.3　文書分類器による検知の精度の評価

文章分類器とテストデータを用いて，個別スレッドの検知精度の評価を行った．結果の混同行列を表2，適合率・再現率・F値，それらのマクロ平均を表3に示す．高リスクなカテゴリ二種について，高い精度で検知ができていることを確認できる．

5.4　算出したリスク指標による診断の評価

テストデータの対象カテゴリについて，中間カテゴリ毎のリスク指標を算出した結果を図1に示す．

定量評価を行う手法が定まっていないため，主観で評価した所，リスク指標が示すような特徴が見られ，有望と考えられた．例として，中間カテゴリ「二次元」について,目視で確認をした所，ゲームユーザーや企業社員に対する罵倒を含むスレッドが複数件あり，リスク指標と一致していると見られた。

表2:混同行列

		推定結果		
		政治・宗教等	炎上・アンチ等	対象ｶﾃｺﾞﾘ
正解	政治・宗教等	701	4	30
	炎上・アンチ等	2	286	39
	対象カテゴリ	32	13	2514

表3:適合率・再現率・F値,マクロ平均

	Precision	Recall	F1
政治・宗教等	0.9537	0.9537	0.9537
炎上・アンチ等	0.9439	0.8746	0.9079
対象カテゴリ	0.9733	0.9824	0.9778
マクロ平均	0.9570	0.9369	0.9465

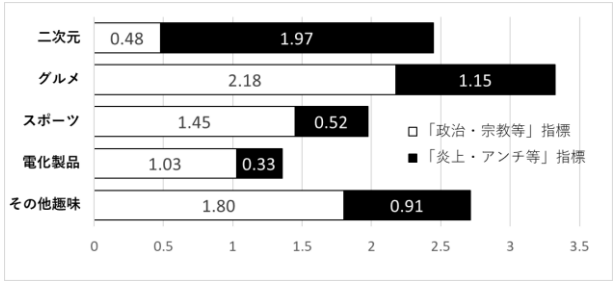

図1:対象カテゴリ(中間カテゴリ毎)のリスク指標

6　おわりに

本稿では，fastTextを用いた文章分類器によって,高リスクな投稿(スレッド)/カテゴリの検知とリスク指標算出による診断を行う手法を提案した．スレッドの検知は高い精度で行うことができると示した.カテゴリの検知とリスク指標も，主観評価では妥当であり，有望であった．今後の課題としては，カテゴリの検知とリスク指標の定量評価を行う手法の検討，高リスクと推定されたスレッドについて，リスクの原因となっているレス(コメント)を特定する手法の検討が挙げられる．

参考文献

[1] Piotr Bojanowski, Edouard Grave,Armand Joulin, Tomas Mikolov. “Enriching Word Vectors with Subword Information”. Facebook AI Researc, 2016.

[2] 大西真輝,澤井裕一郎,駒井雅之,酒井一樹,進藤裕之.ツイート炎上抑制のための包括的システムの構築. 2015年度人工知能学会全国大会（第29回），2015

[3] 竹内幹太,伊東栄典. 文書分類手法による炎上動画検出手法の検討.情報処理学会研究報告2021（B3-3），1-4，2021

[4] 齋藤慎悟，新美礼彦. YouTube Liveでの生配信をきっかけとした炎上の早期発見を行うシステムの提案. 第85回全国大会講演論文集2023（1），545-546，2023