6K-03

マイクロブログ炎上事例の分類と炎上防止アプリへの応用

諸岡 誠也† 関 和広‡
甲南大学 知能情報学部

1 はじめに

Twitter に代表されるマイクロブログが日常的に利用されるようになるのと同時に，自身の発言に対し非難や批判が殺到する「炎上」の事例も多発している．本研究では，Twitter 上で炎上した過去の事例をその投稿内容によって類型化し，これを基にテキスト分類の手法によって炎上する投稿を検知する枠組みを構築する．また，この枠組みを利用して炎上の可能性がある投稿をあらかじめ検知・警告する炎上防止アプリケーションの開発について報告する．

2 関連研究

炎上リスクに対するシステムの実装として，山下ら [1] は，モニタリングツール「まもっチャオ」の開発を行った．このツールでは，炎上によって意図しない情報が流出する可能性をリスクと捉え，ユーザが設定した NG ワードや個人情報によって危険と判断した書き込みにアラートを送信する機能を備えている．しかしながら，このシステムは書き込み後にしか機能しないため, ユーザがアラートに気づくまでの批判・非難は少なからず発生する可能性がある．

一方，SNS における個人情報の保護の研究として，安藤ら [2] は，複数のサイトで公開された情報を統合することで想定外の個人情報が漏洩し，炎上のような事件に繋がる可能性があることに着目した．そして，複数サイトにおいて公開した情報と，ユーザと友人関係にある他のユーザが公開している情報を統合して，複数のサイト上の自身に関する情報を把握するための支援システムを研究している．しかしながら，このシステムの処理対象は当然ながら複数の SNS サイトを使用しているユーザに限られる．

文書分類を用いた炎上防止に関する研究として，岩崎ら [3] は，「ある話題について自分の評価を強引に押し通すような投稿」が炎上しやすいことに着目し，このような文章を対象に炎上を未然に防ぐ事を目的とした研究を行った．しかしながら，SNS に投稿される文書は多種多様であり，他の種類の投稿についての有効性は明らかではない．

本研究では，多様な炎上を検出するため，まず過去の多数の炎上事例を人手で類型化する．そして，そこで得られた類型に従って段階的に分類を行う．そして，以上の枠組みを利用して，炎上を事前に防止するウェブアプリケーションを開発する．

3 炎上事例の類型化

一言で炎上と言っても，すべての事例が同じ条件で炎上しているわけではない．例えば，発言者の立場が異なる場合，必ずしも同じ発言が炎上するとは限らない．そこで本研究では，炎上事例をより正確に分類するために，このような発言者の立場に着目した類型化を行った．類型化の作業のため，関連研究 [1] で収集された炎上事例 80 件から添付画像によってのみ炎上した事例 11 件を除いたものに加え，新たに 20 件の炎上事例をウェブ上から人手で収集し，計 89 件の事例を対象に分析を行った．まず，これらの炎上事例を「意見を与える側」と「意見を受ける側」の 2 つの立場に着目してグループ化した．

次に，それぞれのグループごとに炎上事例の詳細を観察したところ,「意見を受ける側」の炎上事例の中でも,「犯罪ではないが問題がある行動の暴露」によって炎上したものと,「犯罪の暴露」によって炎上したものでは，発信者の立場が異なる事が分かった．これは，何らかの組織に所属している事が周知な発言者と個人情報が秘匿されている発言者との違いである．何らかの組織に所属している事が周知な発言者は,「意見を与える側」に類型化された事例にも複数見られ，個人情報が秘匿されている発言者が「過度な暴言」によって炎上しているのに対し,「軽度の悪口」によって炎上している．この事から，何らかの組織に所属している事が周知な発言者は，個人情報が秘匿されている発言者に比べて悪性の低い発言でも炎上してしまう傾向にあることがうかがえる．

以上の考察から，炎上事例は「個人情報が秘匿され

Categolizing flamed microblog messages and its application to flaming prevention
†Seiya Morooka
‡Kazuhiro Seki
Faculty of Intelligence and Informatics, Konan University

ている発言者の他者への誹謗中傷」,「個人情報が秘匿されている発言者の犯罪の暴露」,「何らかの組織に所属している事が周知な発言者による不用意な発言」に類型化できると考えられる．さらに，この類型化を炎上していない事例にも当てはまるよう，各カテゴリを他人や他の物への個人の意見（以下「意見」），自分の近況や変化（以下「自身」），仕事関係のつぶやき（以下「仕事」）と定義する．

4 炎上予測

前節の類型化によって得られたカテゴリを基に，任意の文書（つぶやき）をテキスト分類の手法によって段階的に分類することで炎上予測を行った．具体的には，まず事例の発言者の立場に着目したカテゴリの分類を行い，続いて，発言者の立場における炎上の違いを考慮するために，カテゴリ毎に学習したモデルを用いて炎上するか否かの分類（炎上予測）を行った．

分類器としては，カテゴリ分類および炎上予測の 2 段階とも，サポートベクターマシン（SVM）[4] を用いた．なお，SVM には線形カーネルを用い，コストパラメタ C はそれぞれの分類器ごとに最適な値に設定した．学習に用いる各文書は，形態素解析の後，サ変接続・数値以外の名詞，非自立動詞以外の動詞，副詞，感動詞のみを抽出・原型化し，さらに「笑」・名詞以外の原型にできなかった単語を除き，TFIDF 法によってベクトル化した．実験データは，3 節で分析に用いた 89 件の炎上事例に加え，ウェブ上から人手で収集した 160 件の非炎上事例を用いた．評価尺度としては，適合率，再現率，F 値を用いた．leave-one-out 交差検定でグリッドサーチを行い，カーネルは線形カーネル，コストパラメタ C はカテゴリ分類で 100 とし,「意見」,「自身」,「仕事」カテゴリでの炎上予測でそれぞれ 10, 100, 1 とした．なお，炎上予測の分類の際は，カテゴリによって正事例数と負事例数の差が大きいことから，ダウンサンプリングによって事例数を調整した．比較対象には，カテゴリ分けを行わない通常の SVM 分類器（以降では「1 段階 SVM」と呼ぶ）を用い，同様に最適なカーネル，パラメタ値を設定した（RBF カーネル，$C = 1000$, $\gamma = 0.0001$）．

表 1 に分類の結果を示す．この結果から，カテゴリ分けによる 2 段階の分類を行った場合には，カテゴリ分けを行わない場合よりも高い精度が得られることが確認できた．

表 1: 分類結果

分類	炎上			非炎上		
	適合率	再現率	F 値	適合率	再現率	F 値
1 段階 SVM	0.494	0.483	0.488	0.551	0.727	0.711
2 段階 SVM	0.715	0.706	0.594	0.769	0.831	0.799

5 炎上防止アプリ

前節で学習したモデルを利用し，主にスマートフォン等の携帯端末での使用を想定した炎上防止アプリ「ついふる」を任意のブラウザで動作するウェブアプリケーションとして開発した．本アプリでは，4 節での処理と同様に，ユーザが入力した文章に対し形態素解析・ベクトル化を行ったのち，2 段階の分類による炎上判定を行う．また，Twitter API をラッピングした Python モジュールである Tweepy を用い，本アプリから Twitter への投稿も行う．以上の実装により，本アプリでは，ユーザがつぶやきを入力したのちに，分類結果とカテゴリを即座に確認でき，炎上すると判定された場合には警告が発せられ，炎上の危険性を低減することができる．

6 おわりに

本研究では，マイクロブログにおける炎上を防止するために，炎上事例を 3 つのカテゴリに類型化し，それぞれのカテゴリ毎に炎上予測を行う炎上防止アプリ「ついふる」を開発した．今後の課題として，より高精度な予測を行うため，ウェブ上から事例を自動で収集する枠組を構築していきたい．

謝辞

本研究の一部は，私立大学等経常費補助金特別補助「大学間連携等による共同研究」によるである．また，Twitter の炎上事例については，東京工業高等専門学校の山下景弘氏から提供を受けた．

参考文献

[1] 山下晃弘，中村拓哉，川村秀憲，鈴木恵二 “SNS における炎上リスク分析と対策システムの開発” 人工知能学会知識ベースシステム研究会資料, Vol. 103, pp. 19-24, 2014.

[2] 安藤寿英，中村健二，小柳滋 “複数 SNS サイトにおける発信情報分析による個人特定の可能性の検証” 情報科学技術フォーラム講演論文集, Vol. 11, pp. 341-342. 2012.

[3] 岩崎祐貴，折原良平，清雄一，中川博之，田原康之 “CGM における炎上の分析とその応用” 人工知能学会論文誌, Vol. 30, No. 1, pp. 152-160, 2015.

[4] Vladimir N. Vapnik, The Nature of Statistical Learning Theory. Springer, 1995.