研究指導 中澤 真 教授

Twitterにおける炎上表現辞書の自動生成
－単語の共起関係に着目して－

丸山 泰智

1. はじめに

近年，TwitterやFacebookなどのSNSで不用意な投稿が原因となって，投稿者である個人や法人のユーザが非難に晒される「炎上」が社会問題となっている[1]．若者がコンビニエンスストアや飲食店で迷惑行為を働き，その写真を投稿したり，官僚や政治家が暴言を吐いたりすることなどが例として挙げられ，年代や職業を問わず多くの人が炎上している[2]．さらに，総務省が2013年に実施したアンケートでは炎上のきっかけとなったSNSとしてTwitterが最も多いという結果が出ている[1]．そのため，数あるSNSの中でも特にTwitterでの炎上発生を抑止することが必要とされている．

そこで，ツイートが炎上する危険性があるかどうかを予測するシステムの開発が行われている[3][4][5]．これらは炎上を予測するためにツイート本文を自然言語処理分析しているものが多い．学習用データ[1]としてのツイートから，炎上に関係する特徴的な単語をあらかじめ抽出し，判定したいツイートにそれらの単語がどの程度含まれているかによって，ツイートが炎上するか否かを予測している．この学習用データであるツイートから抽出される単語の集合が辞書である．この辞書は自動で生成されるものと手動で生成されるものの2種類に大別される．

辞書を手作業で作成した場合は，比較的高い判定精度を達成しているが，辞書の作成にかなりの労力が必要になるという問題がある[5]．一方，辞書を自動生成する場合，作業の効率化を図れるが，現状では判定精度が低いという問題がある[3][4]．自動生成の手法では，炎上ツイートと非炎上ツイートに出現する単語の集合から，炎上ツイートのみに出現する単語を抽出して判定することを試みているが，実際にはそのような単語が多くないため十分な精度が得られないと考えられる．

そこで本研究では，辞書の自動生成をしつつ，ある程度の判定精度を達成するために，単語単体の辞書だけでなく，連なる二つの語など共起関係に着目した辞書を追加することで，この問題の解決を図る．

2. Twitterにおける炎上警告システム

2.1 炎上警告の概要

炎上とはSNSなどで発信した情報に対して，多くの批判が寄せられる現象である[6]．この現象は，犯罪行為の助長や個人，企業への風評被害をもたらすため[3]，炎上の元となる投稿をツイートする前にユーザに警告する必要がある．

そのため，投稿前にその内容に不適切な部分がないことを，システム的にチェックする機能が求められる．これを実現するために，炎上しやすい特定の単語が投稿内容に含まれているか否かによって炎上の危険性を予測し，ユーザに警告する研究が行われている．例えば，川上ら[3]は炎上する危険性があるテキストに対して投稿前に警告するシステムを開発している．こうしたシステムによって，炎上の危険性を事前にユーザへ通知し，不本意な炎上を防ぐことが可能になる．

2.2 炎上警告における辞書

前節で述べたような炎上警告システムでは，炎上ツイートに含まれる特徴的な単語である特徴語を見つけることが重要となる．炎上の可能性を判定するのに有用な特徴語は，炎上ツイートに頻出し，問題のないツイートには現れない傾向にあるものが望ましい．こうして選ばれた特徴語の集合が辞書であり，本研究ではこれを炎上表現辞書と呼ぶ．

炎上表現辞書の作り方には炎上ツイートには出現し，一般ツイートには出現しない名詞と動詞を抽出する手法[4]がある．また，辞書をタイプ別に複数用意する研究[3][5]もある．例えば，既存の研究で生成された新語辞書，手作業によって生成したネットスラング辞書を組み合わせて単語の類似性から炎上を判定する手法[3]がある．別のものとして，主語と考えられる単語を集めた辞書，主語を否定する単語を集めた辞書，主語を強調する単語を集めた辞書に分け，それぞれの辞書に含まれる単語の組み合わせによって炎上を判定する手法[5]もある．

こうして構成された炎上表現辞書を用いて炎上ツイートであるか否かの判定がなされている．武本[4]は炎上表現辞書を半自動で生成して炎上ツイート80件と一般ツイート160件で炎上判定を行ったが，炎上ツイートを正しく判定できたものは5割程度と低い精

[1] パターンや傾向を掴むために用いるデータ

度となった．複数の辞書を自動生成している川上ら[3] でも，精度が低い結果となっている．一方，手動で辞書を作成した青柳ら[5]の手法では，炎上ツイートの判定精度として約7割を達成している．

2.3 炎上表現辞書を用いた研究の現状と課題

武本[4]や川上ら[3]の研究は辞書を自動で生成しているが精度は低く，青柳ら[5]は人手で辞書を生成しているため精度は高いが，生成するために必要な時間や手間が多くかかってしまうという問題があった．

そこで，本研究では炎上表現辞書を自動で生成させることに加えて，精度の低下を防ぐための手法を提案する．まず，どのような単語を辞書に登録するべきか予備調査を実施したが，炎上ツイートに出現していて，一般ツイートに出現していない単語はほとんど存在しないことが明らかになった．武本[4]の手法ではこうした単語を利用しているため，辞書に登録できる単語が極端に少なくなり十分な精度を得られていないと考えられる．そこで，単語単体の辞書に加えて，共起関係[2]にある単語の対からなる辞書を生成することを考える．単語単体では，炎上ツイートと一般ツイートの両方に出現してしまっていても，2語が同時に出現する共起関係の場合は，どちらかに偏って出現している可能性が高まり，判定のための特徴語として有用であることが期待できる．

以上の理由から，本研究で自動生成する辞書は精度を高めるために，単語単体の炎上表現辞書と単語の共起関係に着目した炎上表現辞書を生成し併用する．

3. ツイートの抽出と辞書の自動生成

3.1 ツイートの抽出

まず，分析に使用する炎上ツイートと一般ツイートを抽出する．なお，画像や動画を含むツイートでは，文章そのものには投稿者の主張が含まれないことが多いため，今回は文字のみのツイートに限定する．

次に，以下の二つの条件を両方満たすものを炎上ツイートと定義する．

① リプライまたは引用リツイートで「通報しました」という文節が含まれている．

② ツイートの内容を明らかに否定している批判リプライが5件以上寄せられている．なお，ここでの，批判リプライとはツイートの内容を明らかに否定しているリプライとしている．

①を用いた理由は，炎上ツイートに対して「通報しました」というリプライが送られた際に，実際に通報している場合と危機感を煽るためにこのように発言している場合が考えられ，炎上の前段階と捉えられるからである．しかし，この言葉は仲間内の冗談で使用されることもあるため，この条件だけで直ちに炎上ツイートであると判断することはできない．そこで，一定数の人が不適切だと感じて批判的なリプライを投稿しているかという②の条件を加えることにより，単なる冗談で用いている場合のツイートが除外されるようにした．5件以上の批判的なリツイートがあれば，個人のアカウントとしては十分に炎上状態と考えられるからである．

次に，ツイートの抽出方法について述べる．炎上ツイートの抽出は「通報しました」というキーワードでTwitter内を検索してヒットしたもの中から，批判リプライが5件以上あるツイートをランダムに抽出した．通常の一般ツイートの抽出方法は以下の通りである．まず，フォロワーが10000人以上のアカウントを抽出する．さらにそのアカウントのフォロワーの中から，7日以内にツイートしていて，かつフォロワーが100人以上のものを選び出した．そのアカウントから，1人につき1ツイートをランダムに抽出した．このようにした理由は，ツイートの内容に偏りが生じないようにし，Twitterを頻繁に利用していることに加えて発言に影響力が生じるアカウントを抽出するためである．また，内容に関係なくランダムに抽出したツイートに加え，炎上ツイートとの判定が難しい政治や差別の内容に関わる一般ツイートを別途抽出した．炎上する発言には政治や差別に関するものが多く，これらの内容に関連する一般ツイートと比較すると，類似した単語が出現しているため判定がより困難になる．このような一般ツイートを炎上と誤判定しないためにも，分析対象のツイートとして加えておく必要があると考えた．政治・差別関連の一般ツイートは，政治や差別をハッシュタグとしているツイートを検索した．なお，ハッシュタグに基づいた検索をしているため，分析の際にはハッシュタグを除外し，ツイート本文のみを対象とした．また，フォロワーが0人のアカウントのツイートは問題となる発言をしても炎上しない可能性が高いため，これらも対象から外した．以上の条件を用いて，炎上ツイート102件，一般ツイート451件（通常201件，政治・差別関連250件）を抽出した．

この抽出したツイートを学習用データと評価用データに分けて用いる．学習用データは辞書を生成するために用いるツイートであり，評価用データは生成された辞書の精度を検証するためのツイートである．ここでは，学習用データとして炎上ツイート52件，一般ツイート401件（通常176件，政治・差別関連225件）を使用し，評価用データには炎上ツイート50件，一般ツイート50件（通常25件，政治・差別関連25件）を用いる．

3.2 辞書の自動生成

先ほど説明した学習用データを用いて，本研究では3つの辞書を自動生成する．1つ目の辞書の内容としては，炎上ツイートに出現する傾向が強い単語で

[2] 自然言語処理における共起とは，任意の文書や文において，ある文字列とある文字列が同時に出現することである

構成されている．残り2つの辞書は，一つのツイート内に出現している二つの単語同士の距離に着目した共起語で構成したものである．

まず，単語単体の辞書を自動生成する．統計解析パッケージに日本語形態素解析器を組み込んだRMeCab[3]を用いて炎上ツイートの学習用データから単語の出現頻度および品詞を求めた．その中から，単語単体で意味を持つ名詞（固有名詞，一般，サ変接続）を対象として，辞書に登録する単語を抽出する．表 1はこの分析によって得られた頻度上位10単語である．

表 1 炎上ツイート（名詞）の出現頻度上位 10 単語

順位	単語	品詞	品詞細分類	出現頻度
1	日本	名詞	固有名詞	18
2	国	名詞	一般	8
3	安倍	名詞	固有名詞	7
4	韓国	名詞	固有名詞	7
5	憲法	名詞	一般	6
6	差別	名詞	サ変接続	5
7	批判	名詞	サ変接続	5
8	自分	名詞	一般	5
9	社会	名詞	一般	5
10	人間	名詞	一般	5

また，再現率を高めるため，出現頻度が低い名詞でも，一般ツイートには出現せず，炎上ツイートの出現頻度が2以上のものも辞書に登録する．これらの単語により構成された辞書を本研究では炎上語辞書と呼ぶ．

次に，共起を用いた辞書を自動生成する．共起を用いた辞書は二種類生成する．まず，N-gramを用いた辞書の説明を行う．N-gramとは，文字や形態素，品詞などの連なりのことである．Nは任意の数字で，いくつの要素の連なりであるかを表す[7]．本研究では2単語の共起関係のみに着目しているため，連続して出現する2対の単語である2-gramを用いることにした．2-gramを登録する炎上2-gram辞書を生成するために，RMeCabを用いて炎上ツイート，一般ツイートのそれぞれにおける2-gramの出現頻度を求め，炎上ツイートでの出現頻度が2以上かつ一般ツイートには出現していない単語の対をこの辞書に登録する．ここでは品詞を特に指定しない．品詞を限定せずに辞書を生成することで，助詞や記号などように一般ツイート・炎上ツイートの両方で出現する単語が，どの単語とともに出現した場合には炎上しやすいのかわかり，精度の向上が期待できる．生成された炎上2-gram辞書の一部を表 2に示す．

表 2 炎上 2-gram 辞書（抜粋）

1語目	2語目	出現頻度
の	か	25
！	！	24
」	と	20
だ	う	19
障害	者	17
方	が	16
し	、	15
差別	する	14
、	「	12
。	これ	11

2-gramでは連なる2つの単語の共起関係に着目したが，今度は隣接せずに近隣に出現している2対の単語の共起関係について考える．ここで2対の単語の近さは，間に挟まる単語の数によって決まる．この単語間の距離を4以下とした場合の共起語の対を抽出する．距離を4に設定した理由は，3や5と比較して共起の精度が向上したためである．この共起関係の頻度が高いものを抽出する際には，基点となる1語をあらかじめ決めておく必要がある．この基点となる語は炎上ツイートにおいて出現頻度が6以上となる高頻度のものを選んだ．なお，選ばれた語が一般ツイートに出現していない場合は，単体の炎上語辞書に登録済みであるため除外する．

このように抽出した共起語はすべて炎上共起語辞書の単語として登録する（表 3）．その際，学習用データ上での出現頻度に基づき3つのクラスに分類して登録した．これは，表 3の出現頻度の合計に応じて，3回以上のものをクラス3，2回のものをクラス2，1回のみのものをクラス1とした．

[3] http://taku910.github.io/mecab

表 3 炎上共起語辞書（抜粋）

共起語1	共起語2	出現頻度	出現頻度の合計
ない	騒ぐ	4	5
騒ぐ	ない	1	
ば	騒ぐ	2	5
騒ぐ	ば	3	
て	批判	4	4
批判	て	0	
と	返す	1	4
返す	と	3	
に	騒ぐ	2	4
騒ぐ	に	2	

4. 共起についての有効性の検証

本研究で作成した辞書に基づいて，ツイートが炎上する可能性があるかどうかを判定する手法を以下に示す．

判定対象のツイートに対し，炎上語辞書の単語の出現の有無を変数a，炎上2-gram辞書の単語の有無を変数b，3つに分けて登録した炎上共起語辞書については，それぞれのクラスの単語の出現の有無を変数c3，c2，c1とする．いずれの変数も該当する単語がツイートに出現した場合は1，そうでない場合は0の値をとるものとする．以下が炎上危険度の判定式である．なお，係数の値は様々な試行錯誤の結果，最も精度が高い組み合わせとなるものを選んだ．

$$y = a + b + c_1 + 2c_2 + c_3$$

この判定式を利用し，y>2であれば炎上の危険性があると判定する．ここでは，前節で学習データを用いて自動生成された辞書の有効性を検証するために適合率・再現率を求める．なお，適合率・再現率は以下の式で算出する．

$$適合率 = \frac{炎上と判定したツイートの中で正しかったツイート数}{炎上と判定したツイート数}$$

$$再現率 = \frac{炎上と判定したツイートの中で正しかったツイート数}{炎上ツイートの総数}$$

適合率・再現率を算出した結果を表 4に示す．再現率は先行研究と同程度の7割以上で，適合率は先行研究にわずかに劣る6割程度となったが，判定が難しい政治・差別関連の一般ツイートを含めていることから，精度としては同程度と考えられる．

表 4 適合率・再現率

適合率	再現率
63%	72%

5. むすび

本研究では，辞書の自動生成をしつつ，ある程度の判定精度を達成するために，単語単体の辞書だけでなく，連なる二つの語など共起関係に着目した辞書を追加して，この問題を解決した．辞書の自動生成では単語単体の炎上語辞書に加えて，2-gramと共起語を用いた炎上2-gram辞書，炎上共起語辞書を自動生成し，手動で辞書を作成した先行研究の判定精度と同等の水準に達することができた．

しかし，適合率については改善の余地がある．その理由は，炎上共起語辞書のクラス1の共起語は，炎上ツイートだけでなく一般ツイートにも多く出現していたことで，誤判定が増加してしまった．そのため，特徴語の見直しや，統計的な判別分析などによる精度の改善を図りたい．

参考文献

[1] 総務省，平成 27 年版情報通信白書，http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h27/html/nc242210.html，(参照 2020-2-4).
[2] 日本経済新聞，ネット「炎上投稿」の憂鬱，企業の巨大リスクに，https://www.nikkei.com/article/DGXNASFK1904R_Z10C13A8000000/，(参照 2020-2-4).
[3] 川上幹，彌冨仁，”Twitter への投稿テキストによる炎上警告システムの構築”，ファジィシステムシンポジウム 講演論文集 FSS，vol.32，pp.705-708，2016.
[4] 武本飛鳥，“Twitter における炎上の検知と警告提示手法”，甲南大学，灘本研究室，2016.
[5] 青柳翔，服部哲，”Twitter への擬似犯罪発言抑止におけるリスト組み合わせ方式の提案”，情報処理学会研究報告グループウェアとネットワークサービス(GN)，Vol. 2012-GN-83 pp. 1-7，2012.
[6] 川上量生，ネットが生んだ文化 誰もが表現者の時代，角川学芸出版，2014.
[7] 石田基広，R によるテキストマイニング入門 第 2 版，森北出版，2017.