リプライのポジネガ極性を用いたTwitter炎上の分類手法の提案

渡辺みずほ† 佐藤 哲司††

† 筑波大学情報学群 〒305-8550 茨城県つくば市春日 1-2
†† 筑波大学図書館情報メディア系 〒305-8550 茨城県つくば市春日 1-2
E-mail: †{mizuho19,satoh}@ce.slis.tsukuba.ac.jp

あらまし SNS では，誰もが自身の考えや出来事をネット上に容易に投稿できる．一方で，その容易さから，不特定多数のユーザによる批判的な意見が殺到するネット炎上が多く発生している．炎上を予測し未然に防ごうとする研究はあるが，その精度は十分とは言えず，炎上の偶発的な発生は避けられない．本研究では代表的な SNS である Twitter を対象に，炎上後に必要となる効果的な対処法の判断を容易にすることを目的として，Twitter 炎上を分類する手法を提案する．ブログでの炎上を類型化した研究では，炎上を 3 種類に大別できるとし，それぞれの種類によって効果的な対処法が異なるとしている．このブログでの炎上分類が Twitter でも相当程度適用できるという仮説のもと，リプライツイートが返信先ツイートに対して肯定的か否定的かを表すポジネガ極性を，リプライツイートに付与する．それを基に返信先ツイートの投稿経過日ごとのネガティブ度を求め，経時的な変化の傾向を用いて分類する．人手で炎上と判断したツイートの日ごとのネガティブ度を表す正解データと分類結果を比較して提案手法の有効性を検証した．その結果，日ごとのネガティブ度では適切に分類されないことが分かった．そこで時間ごとのネガティブ度による分類をおこなったところ指定したクラスタ数に分類することができた．

キーワード Twitter リプライツイート 炎上分類

1 はじめに

Twitter[1]や Facebook[2]，Instagram[3]などの SNS の普及に伴い，誰もが気軽に自身の考えや出来事を発信することができるようになった．一方で企業・個人問わず，投稿者の想定を大幅に超えた批判や誹謗中傷が殺到する「ネット炎上」と呼ばれる現象が発生し [1]，社会問題の一つになっている．最近では飲食店やコンビニの従業員が SNS に投稿した動画に批判が殺到し炎上が発生した事例があり．炎上を収束させるために従業員の解雇や謝罪文の掲載などの対応に追いこまれた企業もある．Twitter におけるネット炎上（以下では炎上と略記する）は，ツイートの投稿者が制御できないほど多くのネガティブな内容のリプライが送られることが問題である．言い換えれば炎上の解消のためには投稿者が自身のしたツイートに対するリプライを抑制するための対処をすることが重要である．

炎上に関する関連研究として，Twitter でのツイートへのリプライに対して感情分析を用いて炎上の検出・分析を行うもの [1] や炎上したツイートに含まれる炎上キーワードを抽出し，それに基づき投稿しようとしているツイートが炎上する可能性を判定し警告を行う手法を提案するもの [2] がある．伊地知 [3] によるとブログでの炎上は，反社会的な行為を自慢したり身分を偽って投稿したりするなどの原因で発生する「批判集中型」，あやふやなことを断言するなどの原因で発生する「議論過熱型」，そしてありとあらゆる原因が考えられる「荒らし型」の 3 種類に分類できるとされる．

このように炎上に関連する研究では，炎上ツイートの文面やそのツイートへのリプライから炎上するかどうかをある程度予測できると考えられる．また，先行研究では炎上ツイートに対するポジティブ（肯定的）なリプライ・ネガティブ（否定的）なリプライの数から炎上の特徴を捉えることはできていない．さらに伊地知によって 3 種類に分類された炎上はブログを対象としたものであり，SNS についても同様に分類できるかという主旨の研究は行われていない．

2 関連研究

本研究は，伊地知 [3] のブログでの炎上は批判集中型・議論過熱型・荒らし型の 3 種類に分類できるという根拠に基づいてツイートの炎上を分類することを目指している．

中川 [4] も同様に書き込みの内容によって分類しており，「義憤型」「いじめ型＆失望型」「便乗＆祭り型」「不満＆怒り吐き出し型」「嫉妬型」「頭を良く見せたい型」の 6 つに分類している．

伊地知による分類は炎上そのものに着目したものであり，中川による分類は書き込みの内容に着目したものである．本研究で炎上を分類することにより，最終的には炎上の抑制につながることを期待しているため，分類ごとの対処法を示している伊地知による分類を採用した．

また炎上そのものではなく，炎上に加担する人たちに着目した研究として，横田ら [5] は炎上の検知を目的として，Twitter ユーザを分類した．山口 [6] は同じく炎上加担者に着目し，実

1 : https://twitter.com/
2 : https://www.facebook.com/
3 : https://www.instagram.com/

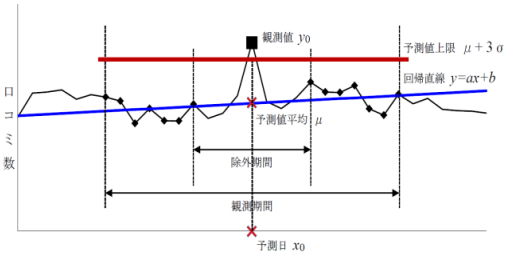

図 1　炎上検知アルゴリズムの概要　　出典：大曾根匡 (2016) p.3

証分析を行い炎上加担者の属性や割合を明らかにした．

本研究ではリプライツイートが元ツイートに肯定的か否定的かを示す極性を，日本語評価極性辞書 [7] をもとに辞書を作成し算出している．高橋 [1] らも同様に日本語評価極性辞書をもとにした極性値辞書を作成し，リプライツイートとの適合度から炎上の検出を試みている．その手法では顔文字に極性値を与えたり，「とても」「～ない」など，文章中の他の単語に影響を与える単語に値（影響値）を付与したりするなどして，極性値辞書を拡張した上で，リプライツイートの極性を算出する手法を提案している．

大曾根 [8] らは炎上の規模をクチコミ数の時間推移データによって定量化し，ネット炎上を検出するアルゴリズムを提案している．このアルゴリズムは図 1 のように，ある一定の期間を除外したデータから回帰直線を導出し，除外した期間内のある日の実観測値が回帰直線による推定値より大きく外れているときにその日を話題拡散日とし，その日におけるクチコミの内容がネガティブなものである場合，その日を炎上日とするものである．クチコミの内容がネガティブなものかどうかは人手で判断する必要があるが，本研究はある一つのツイートに対する日ごとのポジティブリプライ数とネガティブリプライ数の比率をもとに，python の時系列データ分類ツール tslearn を用いて炎上分類を行うため，リプライツイートの内容を逐一人手で判断する必要がない．

このように炎上予測や炎上の検出を行う研究はあるが，情報学的観点からすでに発生してしまった炎上を分類することで，どんな対処をするべきかという判断を容易にすることを目的とした研究は未だ行われていない．

3　炎上ツイートを分類する手法の提案

本研究では，リプライツイートのに基づき炎上ツイートを分類する手法を提案する．提案手法の概要を図 2 に示す．まず日本語評価極性辞書 [9] からポジネガ辞書を作成する処理について 3.1 で説明する．次にツイート・リプライ集合とポジネガ辞書との照合や，ポジティブリプライ・ネガティブリプライの比率により，炎上ツイート・リプライ集合の抽出を行う処理について 3.2 で説明する．なおリプライツイートはテキストが「@[リプライ先のアカウント ID]」から始まる特徴があるため，それを利用しリプライツイートを判別する．続いて著者が炎上と判断した 4 件（内訳は批判集中型 1 件，議論過熱型 1 件，荒らし型 2 件）のツイートと，4 件それぞれに対するリプライから正解となるデータを作成する処理について 3.5 で説明する．最後に炎上ツイート・リプライ集合における日ごとのネガティブ度を時系列分類ツール tslearn を用いて複数のクラスに分類する処理について 3.4 で説明する．

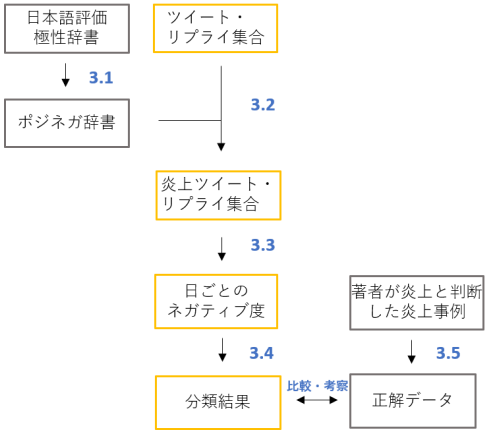

図 2　提案手法の概要

3.1　ポジネガ辞書の作成

リプライツイートがポジティブなものかネガティブなものかを判定するために必要となるポジネガ辞書を，日本語評価極性辞書 [7] を用いて作成する．日本語評価極性辞書とは日本語の用言・名詞ごとにその語が一般的に良い印象を持つか悪い印象を持つか中立の印象を持つかを表すものである．

手順としては，まず日本語評価極性辞書に含まれる単語のうち，評価極性が中立のものを取り除く．そして単語をキーとし，+1, −1 のどちらかを値とするポジネガ辞書を作成する．このときポジティブな単語には+1, ネガティブな単語には −1 を値として付与する．さらに研究する過程で得られた知見として，炎上ツイートのリプライツイートには「てめえ」「アホ」などの罵倒語が多く含まれるため，罵詈雑言辞典 [9] を参考にポジネガ辞書に値を −1 として単語を追加する．また一般的に挨拶は親しい者の間で交わされると考えられるため，「おはよう」や「こんにちは」などの単語は値を +1 としてポジネガ辞書に追加する．

3.2　炎上ツイート・リプライ集合の抽出

3.2.1　リプライツイートのポジネガ判定

続いてリプライツイートの本文を対象に MeCab を用いて形態素解析を行い，名詞や用言の原形を抽出する．辞書には macab-ipadic-NEologd を用いる．さらに式 (1) のように抽出した単語と作成したポジネガ辞書に含まれる単語を照合し，適合した単語の値を順次足していく．最終的に計算された値をリプライツイートの極性値 $p(tweet)$ とする．極性値 $p(tweet)$ が 0 を超えるリプライをポジティブリプライ，0 未満のものをネガティブリプライ, 0 のものをニュートラルリプライと判定する．

$$p(tweet) = \begin{cases} \text{ポジティブ } (+1) & if \quad \sum_{i=0} p(word_i) > 0 \\ \text{ネガティブ } (+1) & if \quad \sum_{i=0} p(word_i) < 0 \\ \text{ニュートラル } (+1) & if \quad \sum_{i=0} p(word_i) = 0 \end{cases} \tag{1}$$

3.2.2 炎上ツイートの抽出

元ツイートの投稿日を起点とし，i 日後に投稿されたポジティブリプライの数を P_{p_i}，ネガティブリプライの数を Pn_i，ニュートラルリプライの数を Pe_i とする．元ツイートへのリプライツイートの比率が以下の式を満たし，かつリプライツイートが 50 以上存在するツイートを，炎上ツイートとみなす．

$$\frac{\sum_{i=0} P_{n_i}(tweet)}{\sum_{i=0} P_{p_i}(tweet) + \sum_{i=0} P_{n_i}(tweet) + \sum_{i=0} P_{e_i}(tweet)} \geqq \frac{1}{3} \quad (0 \leqq i \leqq 6) \tag{2}$$

式 (2) はネガティブリプライ数がリプライ数全体の 1/3 以上 [1] になるツイートと，それに対するリプライを，炎上ツイート・リプライ集合として抽出することを示す．

3.3 日ごとのネガティブ度計算

分類の下準備として，3.2.2 で抽出した上ツイートを対象に，日ごとのネガティブ度を計算する．

炎上ツイートの投稿日を起点とし，期間 i に投稿されたポジティブリプライ数を P_i，ネガティブリプライ数を N_i とすると，期間 i に投稿されたツイートから算出されるその期間のネガティブ度 P_{n_i} は，ツイートの極性値 $P(tweet)$ を用いて，次の式で与えられる．

$$\begin{aligned} P_{n_i} &= \frac{N_i}{P_i + N_i} \\ &= \frac{\sum_{i=0} P_{n_i}(tweet)}{\sum_{i=0} P_{p_i}(tweet) + \sum_{i=0} P_{n_i}(tweet)} \end{aligned} \tag{3}$$

(3) は，期間ごとのネガティブリプライ数を，同期間でのネガティブリプライ数とポジティブリプライ数の和で割った値を示しており，本研究ではこの値をその期間の「ネガティブ度」としている．

3.4 炎上ツイートの分類

炎上ツイートを，日ごとのネガティブ度をもとに，時系列データ分類ツール tslearn [10] を用いてクラスタリングをする．分類するクラスタ数についてはエルボー法を用いて最適な数を求める．なお，クラスタリングを行う際に，相互相関を計算するため，ネガティブ度の平均を 0，標準偏差を 1.0 にスケーリングし，データを正規化する．

図 3 に想定されるポジネガリプライ数とネガティブ度の分類ごとの傾向として，批判集中型と議論過熱型の場合を示す．批判集中型の炎上は，ネガティブリプライ数が日ごとに増加する一方で，ポジティブリプライ数は減っていくと想定される．議論過熱型の炎上は，ネガティブリプライ数もポジティブリプライ数も同程度に増減すると想定される．

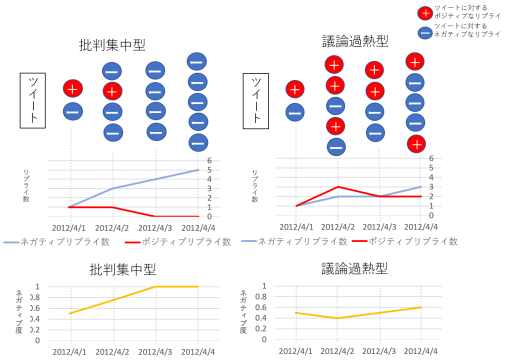

図 3 ポジネガリプライ数の分類概念

この傾向があるとすると，図 3 のように批判集中型の炎上のネガティブ度は徐々に 1.0 に近づき，批判集中型の炎上のネガティブ度は 0.5 に近づく．

3.5 正解データの作成

人手により炎上と判断した 4 つの事例（批判集中型 1 件，議論過熱型 1 件，荒らし型 2 件）をもとに，正解となるデータを作成する．分類対象としたデータと同様に，リプライツイートがポジティブリプライかネガティブリプライかを判定し，それを利用し元ツイートの日ごとのネガティブ度を求める．分類結果と比較するため，ネガティブ度を平均 0.0，標準偏差 1.0 になるようスケーリングし，可視化する．

4 実験と評価

4.1 データセット

本研究では，独自に収集した日本語ツイートから 2012/4/1 ～ 2012/4/7, 2012/7/1 ～ 2012/7/7, 2012/10/11 ～ 2012/10/17, 2013/1/25 ～ 2013/1/31 の期間に投稿・公開されたツイート 1,195,936,178 件を実験対象のツイート集合とする．

4.2 ポジネガ辞書の作成

日本語評価極性辞書 [9] をもとにして単語に対し +1 か −1 の値を付与したポジネガ辞書を作成した．日本語評価極性辞書には用言編と名詞編があり，用言編には 5,187 個の用言にポジティブ・ネガティブ，名詞編には 8,038 個の名詞にポジティブ (p)・ネガティブ (n)・ニュートラル (e) のいずれかの評価極性が付与されている．

このうちポジティブな意味の単語には +1，ネガティブな意味の単語には −1 を付与し，表 1 のようなポジネガ辞書を作成する．この際，用言編と名詞編のどちらにも含まれる単語が計 8 個，ポジティブとネガティブ両方の属性を付与されている単語が 1 個あったため，それら 9 単語は重複として除き，計 13,486 単語をこの名詞・用言をポジネガ辞書に登録した [4]．ま

4：名詞編・用言編両方に登録されている単語：セーフティー，賛同．清清，満足，満腹，過ち，患い，ダメ
ポジティブとネガティブ両方の属性を付与されている単語：買い得です

```
照れ　　e　　?がある・高まる（存在・性質）
照れ笑い　　e　　?する（行為）
照れ性　e　　?である・になる（状態）客観
照準　　e　　?がある・高まる（存在・性質）
症　　n　　?である・になる（状態）客観
症候群　n　　?である・になる（状態）客観
症例　　n　　?がある・高まる（存在・性質）
省エネ　p　　?である・になる（状態）客観
省エネルギー　p　　?である・になる（評価・感情）主観
省スペース　p　　?である・になる（評価・感情）主観
省スペース性　p　　?がある・高まる（存在・性質）
称号　　p　　?がある・高まる（存在・性質）
称賛　　p　　?する（行為）他人
称美　　p　　?する（行為）他人
称揚　　p　　?する（行為）他人
称揚・賞揚　　p　　?する（行為）自分
笑　　e　　?である・になる（状態）客観
笑い　　p　　?がある・高まる（存在・性質）
笑い顔　p　　?である・になる（状態）客観
笑い事　p　　?である・になる（評価・感情）主観
笑い声　p　　?がある・高まる（存在・性質）
笑み　　p　　?がある・高まる（存在・性質）
笑顔　　p　　?である・になる（評価・感情）主観
```

図 4　日本語評価極性辞書名詞編（一部）

た名詞編にはニュートラルの評価極性を付与されている名詞があるが，そのような名詞は，後に説明するリプライツイートの極性計算を行う際に影響を与えないため，ポジネガ辞書への登録は行わなかった．

表 1　ポジネガ辞書（一部）

key	value
我が物顔	−1
分からずや	−1
悪あがき	−1
青二才	−1
ありがとう	+1
ありがと	+1
こんにちは	+1
おはよう	+1

さらに罵詈雑言辞典 [9] を参考に, 325 個の罵倒語についても値を −1 として追加した．これは研究する過程で，炎上しているツイートへのリプライには「バカ」や「クズ」などの罵倒語が多く含まれるという知見が得られたためである．

また，「おはよう」や「こんにちは」などの挨拶として使われる単語は，値を +1 として 8 単語をポジネガ辞書に追加した．これは一般的に挨拶は親しい者の間で交わされると考えられるためである．

上記のようにして，ポジティブな意味を持つ単語を 5,420 個，ネガティブな意味を持つ単語を 8,400 個，計 13,820 個の単語をキーとし，それぞれに +1 か −1 の値を付与したポジネガ辞書を作成した．

4.3　炎上ツイート・リプライ集合の抽出

4.3.1　リプライツイートのポジネガ判定

3.1 で作成したポジネガ辞書とリプライツイートとの適合から，リプライツイートの極性値を求め，リプライツイートがリプライ先のツイート（以下では元ツイートと略記する）に対し肯定的なポジティブリプライか，批判的なネガティブリプライかを判定する．

まず，テキストが「@ [リプライ先のユーザ id]」から始まるリプライツイートに対し，MeCab による形態素解析を行った．MeCab は分解した単語ごとに品詞，品詞細分類，活用形，活用型，原形，読み，発音の順で出力する．MeCab の単語分かち書き辞書に NEologd 辞書を利用した．

(1) のように，3.1 で作成したポジネガ辞書のキーと，リプライツイートに対する形態素解析の出力結果のうち原形を適合させ，適合するキーに対応する値を順次足していき，最終的なリプライツイートの極性値を算出する．

極性値が 0 より大きいものをポジティブリプライ，0 より小さいものをネガティブリプライ, 0 のものをニュートラルリプライと判定する．

4.3.2　炎上ツイートの抽出

4.3.1 のポジネガリプライ判定を，対象の期間に投稿されたリプライツイートを対象に行う．元ツイートの投稿日を起点としたときの経過日を統一するため，元ツイートは対象期間の連続する 7 日間のうち最初の日に投稿されたものに限定する．

元ツイートの投稿日を起点とし，i 日後に投稿されたポジティブリプライの数を P_{p_i}, ネガティブリプライの数を P_{n_i}, ニュートラルリプライの数を P_{e_i} とする．元ツイートへのリプライツイートの比率が以下の式を満たすツイートと, それに対するリプライの中で，リプライ数が 50 以上あるツイートを炎上ツイートとみなす．

$$\frac{\sum_{i=0} P_{n_i}(tweet)}{\sum_{i=0} P_{p_i}(tweet) + \sum_{i=0} P_{n_i}(tweet) + \sum_{i=0} P_{e_i}(tweet)} \geqq \frac{1}{3} \quad (0 \leqq i \leqq 6) \tag{4}$$

(4) は，ネガティブリプライ数が全リプライ数の 1/3 以上の元ツイートを炎上ツイートとすることを示す．

4.4　日ごとのネガティブ度計算

4.3.2 で抽出した炎上ツイート・リプライ集合について，(3) にあてはめて経過日ごとの元ツイートへのネガティブ度を求める．

4.5　炎上ツイートの分類

4.5.1　炎上ツイートの条件

2012/4/1 ～ 2012/4/7 までの 7 日間で投稿された 46,784,387 ツイートのうち, 以下の条件を満たす炎上ツイートは 52 件存在する．同様に 2012/7/1 ～ 2012/7/7, 2012/10/11 ～ 2012/10/17, 2013/1/25 ～ 2013/1/31 の期間に投稿されたツイートのうち以下の条件を満たす炎上ツイートはそれぞれ 42 件, 28 件, 52 件存在する．

条件 1　対象期間の連続する 7 日間のうち最初の日に投稿されたツイート

条件 2　ネガティブリプライ数が全リプライ数の 1/3 以上

条件 3　それ自体はリプライツイートではない

条件 4　ツイートの投稿日からの 7 日間でリプライが 50 件以上送られている

条件 1, 条件 2 については既にそれぞれ 4.4, 3.2.2 で説明しているので省略する．

条件 3 について説明する．本研究では炎上ツイートをリプライツイートではないものと定義している．リプライツイートに

対するリプライツイートは，大元のツイート（元ツイート）と論点や立場が変わる可能性が存在するため，本研究では，炎上ツイートそのものはリプライツイートではないと定義する．

続いて条件 4 について説明する．大西ら [11] の研究では，炎上事例を分析する際に，「引用（リツイート）数が 20 以上ある」ことを炎上事例であるための一つの条件としていたが，本研究で扱うデータセットにはリツイート数に関する情報が付与されていないため，リプライ数を基準としている．

2012/4/1 ～ 2012/4/7 の 7 日間で投稿された元ツイートのうち，1 件以上リプライが送られたものは 2,187,200 件存在する．本研究では炎上分類を目的としており，リプライ数が少ないものは炎上ではない．そのためリプライ数が少ない元ツイートは分類対象から除く．

本研究ではリプライが 50 以上あるということを炎上ツイートであるための一つの条件としている．リプライが 1 件以上ある 2,187,200 件の元ツイートのうち，大部分にあたる 2,186,992 件がリプライ数が 50 未満である．

図 5 はリプライ数の多い順に，左からリプライ数の自然対数を示す点を，元ツイート 1 件につき 1 つ描画したものであり，図 6 は 50 リプライ以上送られている元ツイートをリプライ数の多い順に，左からリプライ数を示す点を描画したものである．

図 5, 図 6 からリプライが 1 件以上あるツイートのほとんどが，リプライ数がごく少数であることが分かる．

よって 50 リプライを炎上ツイートの条件の一つとする．

4.5.2 tslearn による炎上ツイートの分類

まず 2012/4/1 ～ 2012/4/7 の期間での炎上ツイートの分類を行う．python の時系列分析のための機械学習ライブラリ tslearn をもちいて，炎上ツイートを分類する．クラスタリング手法として K-Shape を用いる．

使用するデータは表 2 のような形で csv ファイルに格納されており，元ツイートごとに csv ファイルが存在する．elapsed date は元ツイートが投稿された日を起点とした経過日を表し，negative　degree は経過日ごとのネガティブ度を表す．なおリプライが 1 件も送られなかった場合はネガティブ度を 0 とした．

学習用に成形した csv ファイルからデータを読み取り，tslearn のモジュール TimeSeriesScalerMeanVariance を用いて時系列を各次元の平均が 0, 標準偏差が 1.0 になるように正規化し，n_init=10 としてクラスタリングを行う．ここで n_init は指定した回数だけ初期値の異なるクラスター分析を行い，性能が最良のものを出力することを示しており，今回はその値として 10 を指定した．

初めにクラスタ数を 2 としてクラスタリングした結果，図 7 のようになった．図 7 の横軸は元ツイートが投稿されてからの経過日，縦軸は経過日ごとのネガティブ度を正規化した値を示しており，描画されている線一本が 1 件の炎上ツイートを表している．

ここで elbow 法を用いて，最適なクラスタ数を求めた．結果を図 8 に示す．

図 8 から，クラスタ数が 2 と 4 の所で大きく値が下がっているので，クラスタ数が 4 の場合でもクラスタリングを行った．

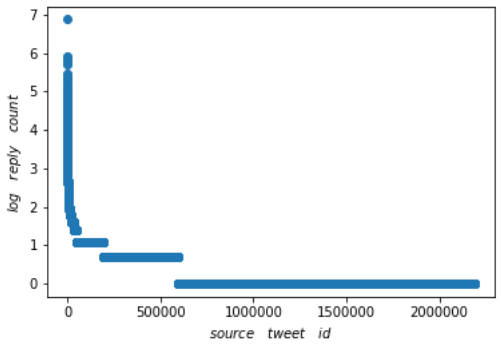

図 5　元ツイートに対するリプライ数　(1 リプライ以上)

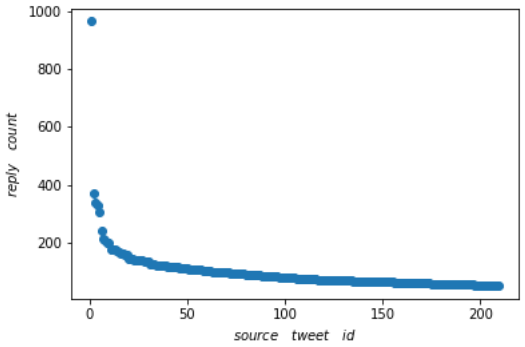

図 6　元ツイートに対するリプライ数（50 リプライ以上）

表 2　学習用に成形した csv ファイルの内容 (一部)

elapsed date	negative degree
0	0.72
1	0.75
2	0.67
3	1.00
4	0.00
5	0.00
6	0.50

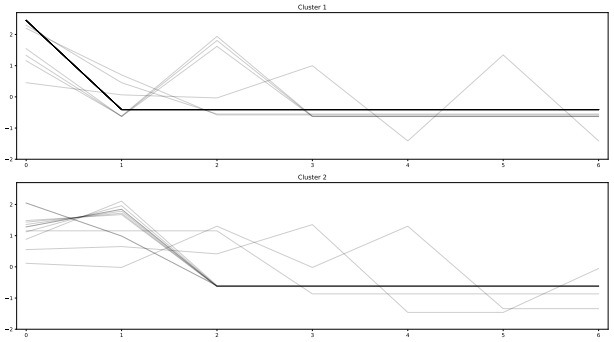

図 7　クラスタ数 2 の場合

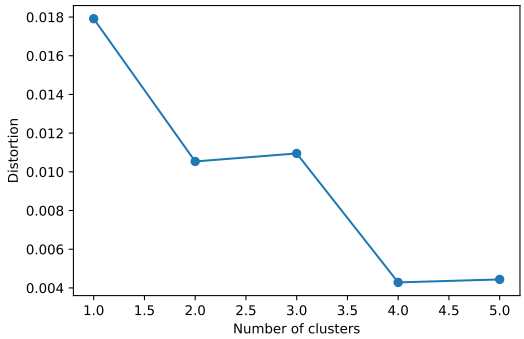

図 8　elbow 法によるクラスタ数の決定

結果を図 9 に示す．

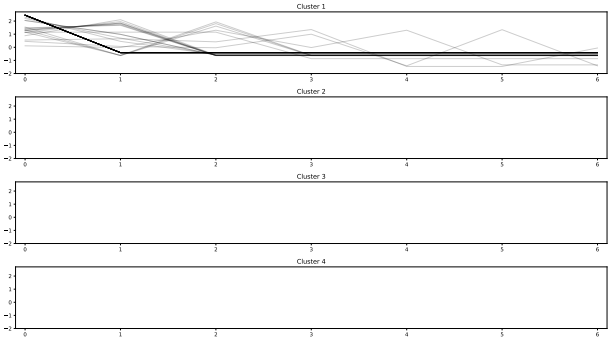

図 9　クラスタ数 4 の場合

しかしクラスタ数が 4 の場合は 1 つのクラスにまとまってしまい，適切にクラスタリングされなかった．

そこで，投稿経過日ごとではなく，投稿経過時間ごとのネガティブ度を算出し，tslearn による分類を行った．

4.5.3　時間ごとのネガティブ度による分類

まず elbow 法によりクラスタ数を決定する．その結果を図 10 に示す．

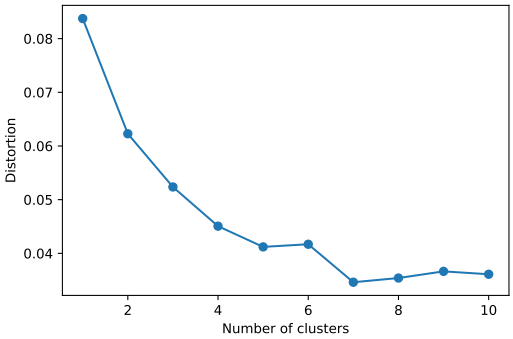

図 10　elbow 法によるクラスタ数の決定

図 10 に着目すると，クラスタ数 4 とクラスタ数 7 の時に値が大きく下がっているため，それぞれのクラスタ数に分類する．その結果をそれぞれ図 11, 図 12 に示す．

また図 7~ 図 12 は 2012/4/1 に投稿された炎上ツイートを

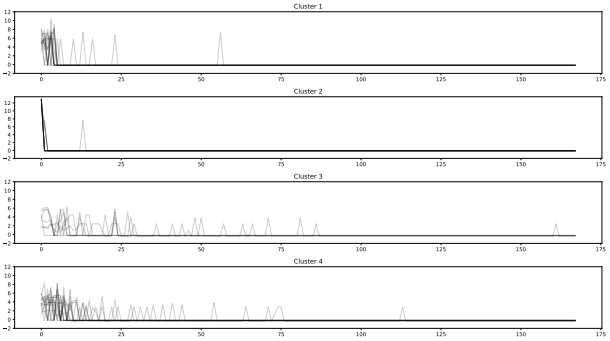

図 11　クラスタ数 4 の場合

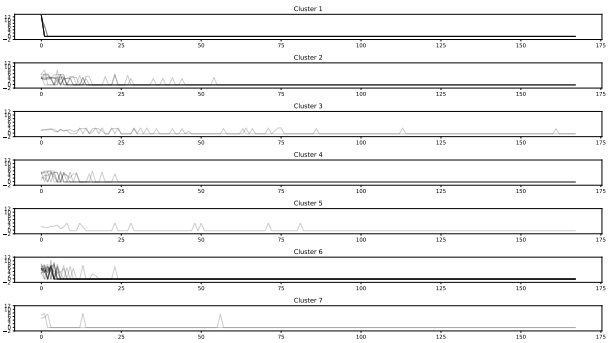

図 12　クラスタ数 7 の場合

対象としたものである．そこで 2013/1/25 に投稿された炎上ツイートを 4 クラスに分類したものを図 13 に示す，

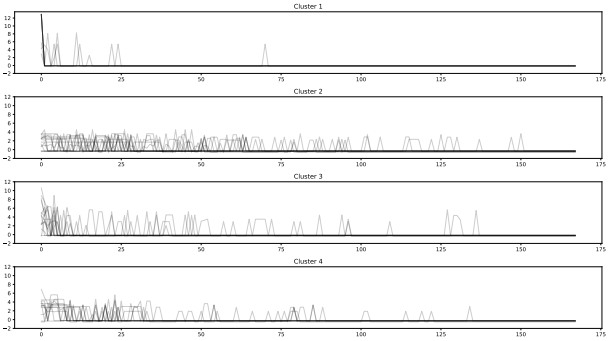

図 13　クラスタ数 4 の場合 (対象期間:2013/1/25 ~ 2013/1/31)

4.6　正解データの作成

筆者が炎上と判断した 4 つの事例から，正解データを作成する．炎上事例 4 件の内訳は, 批判集中型 1 件，議論過熱型 1 件，荒らし型 2 件となっている．

まず批判集中型の炎上事例として，ある一般ユーザによる「これはどう見てもいじめやろ 暴力は良くないよな。どこの学校の生徒なんだろう。」という投稿 [5] を挙げる．この投稿には動画が添付されており，その動画では複数の人物が一人の学生とみられる人物に暴力的な行為をはたらいている．この投稿に対

5：@SoreikeOnePunch による投稿
https://twitter.com/SoreikeOnePunch/status/1140939492242513920
(参照 2019-12-22)

し，「最低」や「捕まって欲しい」「いじめてる方の名前を特定しろ」などのネガティブリプライが多く送られた．

次に議論過熱型の炎上事例として，ある議員による，「上越新幹線東京行き ゆっくりとですが、運転再開しました。ご心配をおかけしました。地震があるたびに、原発は？ と心配になる。本気の原発ゼロへ！」という Twitter への投稿 [6]を挙げる．これは新潟県で発生した地震の直後に投稿されたものである．この投稿に対し，被災者の心配より原発の心配が先なのかというようなネガティブリプライが多く送られた．一方で議員として原発の心配をするのは当然だと投稿者を擁護するポジティブリプライも存在し，議論が交わされている．

次に荒らし型の事例の 1 つ目として，上記の議論過熱型の炎上に発展した投稿の後に，同議員が改めて投稿したツイート [7]を挙げる．その投稿は以下のような内容である．「余震や津波に充分に気をつけてください。震度 6 強を記録した村上市府屋地区は旧山北町。街の中で山形県と入り組んでいるところもあります。避難所に行かれているようで、府屋の皆さんとはまだ連絡が取れていませんが、どうぞお気をつけてください。」この文面自体には非がないが，直前の投稿で炎上したために取り繕っていると捉えられてしまい，「今更遅い」などのネガティブリプライが送られた．

荒らし型の炎上事例の 2 つ目として，あるタレントが投稿したツイート [8]を挙げる．この投稿は「おはおはムキムキ、フサフサドッサーーン♪」という文章に，本人と思われる顔を撮影し加工した画像が添付されている．この投稿者は芸能人であり，投稿者に対して批判的な見方をするファン (アンチ) が多数存在する．そのアンチによって，投稿内容自体には非がないにもかかわらず「うるさい」「気持ち悪い」「朝から嫌なものを見た」などの誹謗中傷とみられる言葉を含むネガティブリプライが多く送られている．

以上 4 件の炎上事例それぞれについて，リプライツイートに対してポジネガ辞書をもとに極性値を求め，ツイートが投稿された日を起点とした，経過日ごと，経過時間ごとのネガティブ度を計算した．

4 件それぞれの経過日ごと，経過時間ごとのネガティブ度を，平均 0.0, 標準偏差 1.0 にスケーリングしたものを図 14，図 15 に示す．グラフは上からある一般ユーザによる批判集中型，ある議員による議論過熱型，同議員による荒らし型，あるタレントによる荒らし型の炎上を表す．

6 : @moriyukogiin による投稿
https://twitter.com/moriyukogiin/status/1140981374041509889 (参照 2019-12-22)
7 : @moriyukogiin による投稿
https://twitter.com/moriyukogiin/status/1141013934859677696 (参照 2019-12-22)
8 : @kurochan96wawa による投稿
https://twitter.com/kurochan96wawa/status/1143636465689063424 (参照 2019-12-22)

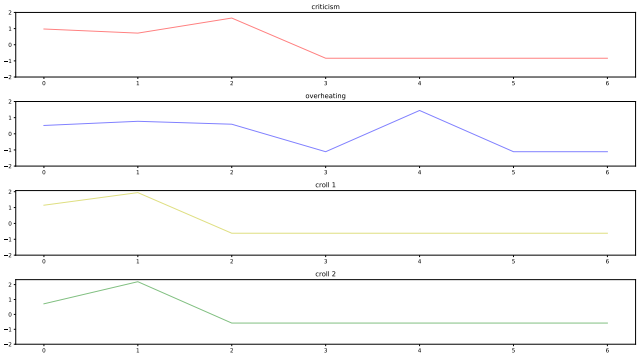

図 14　日ごとのネガティブ度による正解データ

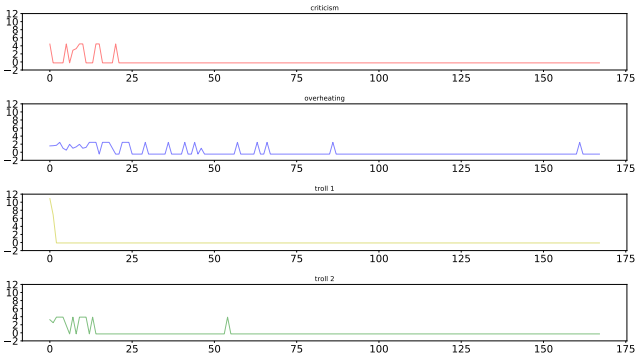

図 15　時間ごとのネガティブ度による正解データ

5 考　　察

5.1 炎上ツイート分類の比較・考察

4. 5. 2 で出力された分類結果と正解データを比較し，本研究の手法の有効性を考察する．なお，ネガティブ度を平均 0.0, 標準偏差 1.0 にスケーリングした値（図 7, 図 9, 図 11〜 図 15 の縦軸にあたる値）を，正規化ネガティブ度と称する．

対象期間を 2012/4/1 ～ 2012/4/7 とした日ごとの正規化ネガティブ度を表す図 7 に着目すると，claster1 では経過日が 0 の時点で正規化ネガティブ度が高いものが多く，元ツイートの投稿日から 1 日後になると正規化ネガティブ度が大幅に低下し，以降その値のまま進む．一方，claster2 では経過日が 0 の時点での正規化ネガティブ度は claster1 同様に高いが，経過日が 1 の時点で正規化ネガティブ度が最大になり，経過日 2 以降は −0.5 付近に収束する．正解データと比較すると，claster2 は図 14 の下 2 つに描画されている，荒らし型のものと近いと考えられる．

図 9 はクラスタリングができていないため，比較・考察を省略する．

時間ごとのネガティブ度による分類では，4 クラス，7 クラスともに指定したクラスタ数に分類することができた．

研究する過程で得られた知見として，「炎上は 1 ～ 2 日で収束し，以降リプライツイートがほとんど送られない」というものがある．経過日ごとのネガティブ度では，炎上の隆盛を適切にとらえることができていない可能性がある．一方で経過日よ

りも単位を細分化した経過時間ごとのネガティブ度による分類では上記の知見を活かして適切に分類できたと考えられる．

また，2012/4/1 はエイプリルフールであるため，この日に投稿されたツイートには「嘘」という単語が多く含まれていた．これがノイズとなり実際には炎上していないツイートを炎上ツイートとして抽出してしまった可能性がある．

さらに，2013/1/25 ～ 2013/1/31 の間に投稿された炎上ツイート 52 件を対象に時間ごとのネガティブ度による 4 クラス分類を行い，炎上ツイートのテキスト本文とリプライツイートを確認したところ，claster1 に「おはよう」や「おやすみ」などの挨拶や短文のものが多く，炎上ツイートとみられるものは存在しなかった．このことから claster1 は非炎上ツイートのグループだといえる．また挨拶を含むツイートへのリプライツイートはツイートが投稿されてから短期間で終息する傾向があるといえる．また claster2 に分類されたツイートの話題として世界情勢や教師の暴力事件に関するものが多く，これらのツイートには比較的長期間にわたりリプライツイートが送られていた．続いて claster3 に分類されたもののうちの一つには，身体的障害を持つ友人に対し Twitter 上で差別発言ともとれるツイートをした芸能人に批判的なリプライツイートが多く送られたものがあった．これは批判集中型の炎上ツイートと考えられる．

最後に，claster4 に分類されたもののなかには電車の中で電話をする女性を撮影した動画を添付し，その女性に対する「きもい」「迷惑」などのコメントを含めて投稿したツイートがある．これに対し投稿者に便乗し被撮影者である女性を罵倒するリプライツイートが多数送られた．一方で，投稿者に対し盗撮ではないかという批判を含むリプライツイートも多数送られており，これは議論過熱型の炎上ツイートと考えられる．

6 おわりに

本研究では，元ツイートに対するリプライツイートの極性を求め，ポジティブリプライ数とネガティブリプライ数をもとに，元ツイートの投稿日からの経過日ごとにネガティブ度を計算し，それを特徴量として，炎上を分類する手法を提案した．

具体的には，日本語評価極性辞書からキーを単語，-1 か $+1$ を値とするポジネガ辞書を作成し，形態素解析したリプライツイートに対して，ポジネガ辞書との適合により付与した極性値により，ポジネガリプライの判定を行った．データセットのうち，複数の条件を満たすツイート・リプライ集合を炎上ツイート・リプライ集合を抽出した．炎上ツイートの投稿日を起点として，経過日ごとのネガティブ度を算出し，python の時系列データ分類ツールとして，k-shape を基にした tslearn を用いてクラスタリングを行った．この分類では，elbow 法を用いて求めた最適なクラスタリング数を適応してクラスタリングを行ったにもかかわらず，有効性を示す結果は得られなかった，

以上のことから，tslearn によるクラスタリングでは，日ごとのネガティブ度は Twitter での炎上を批判集中型・議論過熱型・荒らし型に分類する際に影響を与えないことが明らかになった．

しかし時間ごとのネガティブ度による分類では，指定したクラスタ数に分類することができた．これは時間ごとのネガティブ度が短時間で収束するという炎上の特性に合っているためだと考えられる．

今後の課題としては，tslearn は時系列データ分類ツールであるため，本研究で用いるデータに適していると考えこれを利用したが，時系列間の類似度を測る Dynamic Time Warping(DTW) を用いることでも，炎上分類に役立てられると考える．さらに本研究では人手で炎上と判断した炎上事例 4 件を基に正解データを作成したが，正解データの数が少なく，分類結果と比較できる特徴が少なかったため，正解データを追加することも必要である．

謝　辞

本研究は JSPS 科研費 JP16H02904 の助成を受けたものです．

文　献

[1] 高橋直樹，檜垣泰彦. Twitter における感情分析を用いた炎上の検出と分析. 電子情報通信学会技術研究報告= IEICE technicalreport :信学技報. 2017, vol. 116, no. 488, p. 135-140.
[2] 武本飛鳥. “Twitter における炎上の検知と警告提示手法”. http://www.nadasemi.ii.konan-u.ac.jp/wordpress/wp-content/uploads/2016/12/takemotoS.pdf, (参照 2019-12-20).
[3] 伊地知普一. ブログ炎上 ～Web2.0 時代のリスクとチャンス～. アスキー, 2007, 160p.
[4] 中川淳一郎. ウェブを炎上させるイタい人たち面妖なネット原理主義者の「いなし方」. 宝島社, 2010, 255p.
[5] 横田凌一, 粟屋成崇, 北園淳. 炎上検知のための Twitter ユーザーの分類. システム制御情報学会研究発表講演会講演論文集. 2015, vol. 59, p. 5.
[6] 山口真一. 実証分析による炎上の実態と炎上加担者属性の検証. 情報通信学会誌. 2015, vol. 33, no. 2, p. 53-65.
[7] 小林のぞみ, 松本裕治, 立石健二, 福島俊一. 意見抽出のための評価表現の収集. 自然言語処理. 2005, vol. 12, no. 3, p. 203-222.
[8] 大曽根匡, 福田浩至. クチコミによるネット炎上の定量化の試みとその検証. 情報システム学会 全国大会論文集 第 12 回全国大会・研究発表大会論文集. 2016, vol. 12, p. c22.
[9] 奥山益朗. 罵詈雑言辞典. 東京堂出版, 1996. 348p.
[10] Romain Tavenard, Johann Faouzi, and Gilles Vandewiele. ”tslearn: A machine learningtoolkit dedicated to time-series data”. https://github.com/rtavenar/tslearn. (参照 2019-12-20).
[11] 大西真輝, 澤井裕一郎, 駒井雅之. ツイート炎上抑制のための包括的システムの構築. 人工知能学会全国大会論文集. 2015, vol. 29, p. 1-4.