煽りツイートの自動検出の検討

Examination of Automatic Extraction of Incitement Tweets

松本 典久
Norihisa Matsumoto

岡山大学 太田研究室
Ohta Laboratory, Okayama University

概要 近年，SNS での炎上は対処すべき問題となっている．本研究では，「煽り」や「皮肉」に着目した SNS の炎上防止手法を検討する．そのため，本稿では，Twitter における煽りツイートの自動検出のためにニューラルネットワークによる手法を検討した．作成したモデルに対して，煽りツイートの検出実験により有効性について検討した．

1　はじめに

近年，パソコンやスマートフォンの普及に伴い，SNS はより身近なものとなっている．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，その容易さから不特定多数のユーザによる非難や誹謗，中傷の意見が殺到するネット炎上のリスクも存在する．炎上に関する関連研究として，高橋ら [1] は 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・分析を行った．しかし，「煽り」や「皮肉」，「ネットスラング」への対応が十分ではなかった．これは表面上の意味と真に込められた意味が異なるとき，多くは正しく分析できないためである．また，これらは定型的な内容だけでなく時事を反映した内容も多いことが分析を困難にしている．

本稿ではとりわけ「煽り」に着目し，Twitter 上における煽り表現の含まれた投稿，すなわち「煽りツイート」の自動検出手法を検討する．「煽り」に着目するのは，「煽り」は三者の中で特に発言の対象に対する攻撃的な感情が含まれている可能性が高く，早急な対応が必要だと判断したためである．

本稿ではニューラルネットワークによる教師あり学習が，煽りツイートの検出にどの程度有効なのか検討する．

以下，まず 2 節で煽りツイートついて説明する．3 節では作成したモデルの概要について説明する．4 節では煽りツイートの検出実験について説明する．5 節でまとめる．

2　煽りツイートの分析

「煽りツイート」とは，相手の感情を逆なでる内容のツイートのことを指す．煽り方は多種多様であるが，本稿では数種類に分類する．以下に例を示す．

- 直接的な誹謗中傷や嘲笑
  - 「バカみたいですね」
  - 「そんなんだから友達いないんですよ」
- 皮肉的な煽り
  - (失敗して落ち込んでいる相手に対する) 「おめでとうございます！」
- 見下す，上から目線などマウントを取る煽り
  - 「こんな簡単なこともできないんですか？」
  - 「低学歴のくせに」
  - 「幸せな悩みって知ってる？」
- 相手の言動をわざと悪く受け取る
- ネットスラングなどの定型文を使う

「煽り」には表面上の意味と真に込められた意味が異なるものが混在しており，また時事を反映した「煽り」も存在するため，分析が困難である．

自然言語処理の分析手法の一つに，目的に応じた辞書を作成して分析するものがある．一方，言葉は日々変化するため辞書ベースの分析では新語や造語，俗語の出現が無視できない．また，近年の「コロナ」のように使われ方の変化に対応する必要もある．しかし，辞書の更新を常に手作業で行うのは現実的ではない．つまり，煽りツイートの自動検出には，自動更新される辞書，あるいは辞書を用いない分析の手法がよいと考える．

そのため，本稿ではニューラルネットワークによる手法を検討する．すなわち，辞書を用いない感情分析による「煽り」検出の有効性を検討する．

3　分類モデル

ニューラルネットワークによるツイート分類を行うモデルを図 1 に示す．

これは系列情報を扱うニューラルネットワークである RNN (Recurrent Neural Network) による文章学習とラベル分類を行うモデルである．活性化関数は RNN 層では tanh，出力層では softmax 関数， Optimizer は Adam を使用する．損失関数は出力 $\boldsymbol{y}$ と正解ラベルの Cross Entropy である．出力 $\boldsymbol{y}$ は各次元の要素が，入力がそれぞれのラベルである確率値に等しい 2 次元のベクトルで，各要素は実数範囲 $(0, 1)$ で和が 1.0 となる．学習は誤差逆伝播法を使い，損失関数が最小になるように Adam 法で最適化する．隠れ層は 200 次元，バッチサイズは 64，エポック数は 60 である．各層の役割を以下に示す．

- Embedding 層：単語情報を埋め込み行列に変換．
- RNN 層：入力系列の順方向に順次，再帰的に処理．
- Affine 層：RNN 層からの入力と自身の重み行列との内積を求め，バイアスを足して 1×2 行列に変換．活性化関数として softmax 関数を使用し，出力．

本稿では入力 s_i は入力文を分割し，文頭から順に入力系列とする．ツイートを入力すると本文を単語単位で分割し，単語を ID 化する．ID 化した単語を Embedding 層 (埋め込み層) に入力し，単語情報を行列 x_i に変換して RNN ユニットに入力する．RNN ユニットは前段の RNN の出力した内部

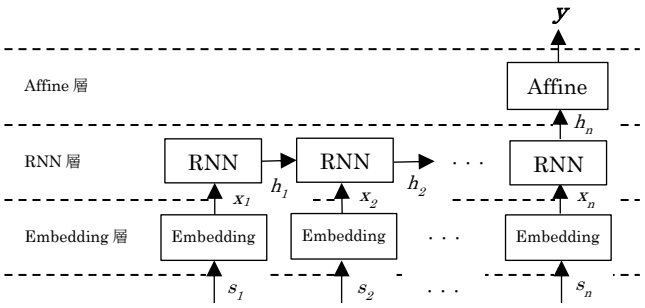

図 1：分類モデル図

状態テンソル h_i と単語情報を合わせて内部状態を計算し，次段の RNN ユニットに出力する．

学習したモデルは，入力された文字列を分類する．分類結果は「煽りである」「煽りではない」の二値である．

4　煽りツイートの検出実験

独自に収集したツイートデータを入力として単純な RNN による「煽りである」「煽りではない」の二値分類が有効かどうかを実験する．

独自に収集した日本語ツイートを対象にモデルを学習，分類を行う．収集対象は画像情報を含まないツイートである．収集したツイートは，表 1 に示すようにツイートの本文からユーザ ID (@以下の英数字列)，ハッシュタグ (#以下の文字列)，URL，画像情報を除去し，本文の文字列のみをデータとして使用する．そのため，本文の文字列が存在しないツイートは除外した．ここで画像情報を除去しているのは，収集の際に誤って画像付きのツイートが混入していた場合への対策である．文字列の分割には，形態素解析で MeCab[1]を用いる．また，MeCab の辞書には mecab-ipadic-NEologd[2]を用いる．

学習データは，GetOldTweet3[3]という Python のライブラリを用いて収集したツイートを使用する．収集した期間は 2020/06/30，2020/07/14，2020/07/18 の 3 日間である．その際，文中に「煽」が含まれているツイートと，「煽」が含まれていないツイートをそれぞれ別に収集した．その後，「煽」が含まれているツイート群から「（煽り），（煽），(煽り)，(煽)」の 4 種類のいずれかが含まれているものを煽りツイートの正解例として抽出し，その後「（煽り），（煽），(煽り)，(煽)」の表現を除去した．

4.1　実験内容

実験は二つ行った．実験 (a) では，正解ラベル付きのデータを用いてモデルの分類精度をテストする．実験 (b) では，ラベルのないツイート群の中から煽りツイートが検出できるのか検証する．

(a)　学習用に煽りツイート，煽りではないツイートをそれぞれ 345 件ずつ使用した．テストに煽りツイート煽りではないツイートをそれぞれ 50 件ずつ使用し，正しく「煽り」かどうかを分類できるか検証した．

表 1：ID，ハッシュタグ，URL 等の除去例

元の文章	@example 今日はいい天気！ #いい予感 http://ex.com pic.twitter.com/Gaz0UReI
除去後の文章	今日はいい天気！

(b)　実験 (a) で使用したすべての煽りツイート，煽りではないツイート各 395 件を学習に使用し，ユーザ ID，URL，ハッシュタグ，画像情報のいずれも含まない無作為に抽出したツイート 10137 件から煽りツイートの抽出を試みた．

4.2　実験結果

まず，表 2 に実験 (a)で行ったツイート分類の結果を示す．表 2 で予測結果が RNN による分類結果である．煽りツイートの適合率は 48/85≒0.56，再現率は 48/50＝0.96，F 値は 0.71 である．

表 3 に実験 (b) で抽出された，煽りと判定されたツイートの例を示す．それぞれのツイートが行われたときの状況を以下に示す．その際，周辺のツイートの内容も加味したうえでこれらのツイートが実際に煽りツイートであるかどうかを人手で判断した．

A)　会話の中での質問に対する回答
- 煽りではない

B)　東京で新型コロナの 1 日の感染者が過去最多となった速報に対するコメントに対するリプライ
- 煽りである

C)　アイドルのパフォーマンスに対する感想
- 煽りではない

D)　なりきりアカウントによる会話劇の一部
- 煽りではない

E)　非公開アカウントとの会話の一部
- 煽りである

F)　ソーシャルゲームで目的のキャラクターを手に入れた人に対するリプライ
- 煽りではない

G)　本の著者に対する好意的な連続ツイートの一部
- 煽りではない

4.3　考察

実験 (a) では煽りツイートのほとんどを「煽りである」と判定している．つまり，再現率は 1.0 に近い．一方，適合率は低いため，この RNN モデルは「煽りである」と判定を下しやすいと判断できる．これは煽りではないツイートは煽りツイートに比べて表現や内容が多様であるため，学習が困難だったためと推測する．したがって，この手法を実用で使うには様々な表現や内容の煽りではないツイートが必要である．

実験 (b) では，表 3 の B，E のように煽りツイートを検出できている．一方で A や D のように煽りとは無関係のもの

[1] https://taku910.github.io/mecab/

[2] https://github.com/neologd/mecab-ipadic-neologd

[3] https://pypi.org/project/GetOldTweets3/

表 2：実験 (a) の分類結果

		予測結果	
		煽りである	煽りではない
正解	煽りである	48	2
	煽りではない	37	13

表 3：実験 (b) で抽出されたツイート例

A	僕も新幹線は長年乗った事がないのでわかりませんけど、確か、車内販売廃止のニュースを見た記憶があります。
B	ウィズコロナのトーキョーですからねぇ。 これが緑の魔女のいうトーキョー大改革 2.0 なんだよなぁ。 小池に投票した奴、ザマァ。って言ったら悪いか？
C	歌声も表情もダンスまで、全てよかったよね 頭の中から離れない
D	あまり、心配させてくれるな……。(お前が居てくれないと)……困る。俺には、まだ(お前が)必要だ。(その身を静かに抱き上げては自身の肩にと乗せてやり)
E	ちなみにこの発動時にこうしてあーしてたら貴方勝ってたかもしれませんよ？あ、あの盤面じゃ無理でしたかｗ
F	わー！！おめでとう(*´ω｀*)
G	クリスチャンだったんだー それなのに柔軟な考えを持ってること、なんだか憧れるなー

も煽りツイートとして判定していることがわかる．また，C，F，G のように煽りとは正反対の好意的なツイートも煽りツイートとして検出している．とりわけ，F と同様に「おめでとう」や「誕生日おめでとう」という内容の複数のツイートが煽りツイートとして判定されていた．これには煽りツイートの一見好意的に見えるツイートが悪影響を及ぼした可能性がある．つまり C，F，G が煽りツイートとして判定されたのは皮肉的な内容だと判断されたためと推測できる．しかし実態は皮肉ではなく文面通りの意味合いだった．つまり，本稿の手法では言葉の一般的な用法と異なる意味を一般的な意味として捉えてしまう危険性が排除できないため，さらなる検討が必要である．

5　まとめ

本稿では，辞書を用いない煽りツイートの自動検出のために RNN による文章分類について検討した．実験結果から提案モデルは B，E のような直接的に感情を逆なでする煽りツイートの検出ができた．一方で C，F，G の判定間違いのように直接的ではない煽りの検出には課題が残る結果となった．原因としては，煽りではないツイートの学習が不十分であることと，ツイート以外の情報を入力していないことが考えられる．したがって，今後の課題としては周辺情報としてツイートの直前，直後のツイートの内容や発言者と煽り対象の関係性，リプライの場合はその前後の会話内容などを組み込んだモデルの設計の検討が挙げられる．

6　参考文献

[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”. 信学技報，vol. 116, no. 488, LOIS2016-86, pp. 135-140, 2017.