Twitter 上のデマ情報の検出

グループ演習 3 班
齋藤和孝, 田嶋脩平, 中村裕
アドバイザー教員 遠藤靖典

1 はじめに

1.1 背景

Twitter とは、世界中の人々が参加するソーシャルネットワーキングサービスの一つである。Twitter では、それぞれのユーザが「ツイート」と呼ばれる最大 140 文字の短文を投稿し、それらをユーザ間で共有することができる。日本でもサービスの開始以来、携帯電話やスマートフォンなどのモバイル端末の普及を背景にユーザ数を急激に増やし、大きな広がりを見せている。Twitter を使うことで、リアルタイムな情報をより多くのユーザに伝播できるため、ソーシャルメディアとしての注目も大きい [1]。実際、多くの企業が新しい広告媒体として利用しているだけでなく、芸能人や有識者も Twitter を利用して情報発信を行っている。例えば 2011 年 3 月 11 日に発生した東日本大震災においても、携帯電話がつながらない状況下での連絡手段として、またよりローカルな情報の発信手段として Twitter は注目を集めた [2]。また、震災に伴い急遽アカウントを開設し情報発信を行う地方自治体も多く見られ、その有用性が確かめられた。

1.2 目的

Twitter の大きな特徴は、既存のメディア [3][4][5] と比べて高速な情報伝播能力を誰でも容易に利用できることである。しかし、共有され伝播する情報の正当性は保証されず、真実とデマ (本論文で扱うデマの詳細な定義については後述する) の判別は困難である。事実、Twitter 上で誤った情報やデマ情報が伝播した事例が数多く存在し、問題視されている [6][7]。誤った情報発信は、情報発信のコストが比較的安価なソーシャルメディアにおいてより顕著にみられ、ソーシャルメディアの持つ有用性を阻害する要因となると考えられる [8]。このような背景を踏まえ、我々は新たなリスクとして認知されつつある Twitter 上のデマ情報の検出を試みる。本研究の目的は、ツイートの信用性を決定づける特徴を抽出し、デマと正しい情報の識別を行うことである。より具体的には、Twitter で実際に投稿されたツイートやツイートを行ったユーザ、ツイートの伝播経路など Twitter 上から得られる情報からデマ情報の特徴を抽出し、評価・考察を行う。

1.3 Twitter の機能

Twitter には、ユーザ間のコミュニケーションを円滑に行うための、様々な特徴的な機能がある。その中から、本研究に関連する 4 つの機能について、その概要を以下に述べる。

- ツイート
  Twitter にメッセージを投稿すること、および投稿されたメッセージのこと。
- リツイート
  他人のツイートを引用して自分のアカウントから発信すること。
- フォロー
  他のユーザのツイートを受信するように登録すること。フォローしているユーザをフォロワーという。
- リプライ
  特定のユーザ宛にメッセージを発信すること。相手のユーザ名の前に@をつけ、「@...」のように記述する。

2 関連研究

Twitter における同種の研究は緒についたばかりで、あまり多くなされていない。本章では、その中でも代表的な 2 つの文献について、その概要を述べる。

2.1 信用性の自動評価に関する研究

Castillo らは、文献 [9] で、ツイートの信用性を自動で評価する技術について述べている。彼らは、「トレンド」に関連する Twitter 上の投稿を分析し、抽出された特徴に基づいて、教師あり学習によりツイートが信用できるか否かを分類している。分析の結果、F 値 86 %で信用性の有無の判断に成功している。ツイートの信用性に影響を与えるものとして、表 1 に示す 15 の特徴を挙げている。

表 1: 信用性に影響を与える特徴

AVG REG AGE	平均登録日数
AVG STAT CNT	平均ツイート数
AVG CNT FOLLOWERS	平均フォロワー数
AVG CNT FRIENDS	平均フォロー数
FR HAS URL	URL を含む頻度
AVG SENT SCORE	平均感情スコア (肯定語の数 - 否定語の数)
FR SENT POS	肯定語の頻度
FR SENT NEG	否定語の頻度
CNT DIST SHORT URLS	短縮 URL 数
SHR MOST FRQ AU	著者が書いた発言の頻度
FR TW USER MENTION	リプライの頻度
FR TW QUEST MARK	？マークの頻度
FR EMOT SMILE	スマイルの絵文字の頻度
FR PRON FIRST	一人称を用いる頻度
MAX LEV SIZE	RT 最大深さ

2.2 自作自演ミームの検出に関する研究

Ratkiewicz らは文献 [10] で、ミームと呼ばれるネットワーク上を伝播する噂・デマの検出に関する考察を行っている。この文献では、Twitter API により与えられたデータの監視、収集、処理を行なう Thuthy システムと呼ばれるシステムを構成している。このシステムの主な機能は、Web 上でユーザが収集されたデータに注釈をつける機能、自作自演ミームを検出する機能、拡散ネットワークの分析を行う機能である。取得したデータを分析した結果、自作自演ミームの拡散ネットワークは、要素のつながりがないミームの起点が異常に多いこと、平均次数が高く星のような形になること、2 要素間のエッジの重みが大きいという特徴を持っていることを明らかにしている。また明らかになった拡散ネットワークの特徴を利用した分類器を作成し、高い精度で自作自演ミームの分類に成功している。

これらの研究はいずれも教師あり学習により分類を行っており、膨大な学習データを必要とする。学習データの作成には多くのユーザからの注釈が不可欠であり、データベースの構築が困難である。そこで、我々は分類にクラスタリング手法を用いることで、膨大なデータを必要とせず、高精度で分類を可能とする手法を提案する。また既存の研究において有用であるとされた特徴が、日本語環境においても有用であるか検討を加える。

3 データ収集

3.1 収集するデータ

本研究では、Twitter に投稿されたツイートを「デマ」と「ニュース」の 2 つに分けて収集している。それぞれの定義を以下に述べる。

- デマ
  本来、デマとは目的を持って意図的に流される嘘を指す [11] [12]。しかし、投稿者の目的に関わらず、嘘の情報がネットワーク上を伝播することに対する社会的影響は大きいと考えられる。そこで本研究では、最初に投稿したユーザの意図に関わらず、内容が事実と異なっている情報は全て「デマ」として定義する。

- ニュース
  社会で実際に起こった出来事や情報を「ニュース」として定義する。

ツイートを収集する際、情報が事実と一致しているかの判断基準として、大手新聞社による報道や公的機関の発表などを利用し、これらの報道された情報は事実としている。

3.2 手法

前述の定義を満たすツイートを web 上で検索し、その中から公式リツイートの多かった 84 ツイート (デマ 30、ニュース 54) を収集した。さらに、これらのツイートをリツイートした 16958 ユーザの登録情報を Twitter API[13] を利用して取得した。本研究で収集した「デマ」、「ニュース」それぞれの例を以下に示す。

「デマ」の例

【拡散希望】千葉市近辺に在住の方！　コスモ石油の爆発により有害物質が雲などに付着し、雨などと一緒に降るので外出の際は傘かカッパなどを持ち歩き、身体が雨に接触しないようにして下さい !!

今年の京大文化祭のテーマは「大切なことはみんな Yahoo!知恵袋が教えてくれた」だそうです

「ニュース」の例

速報:菅首相は衆院予算委員会で高速増殖炉もんじゅについて廃炉を含め検討すべきだとの認識を示した。http://bit.ly/17n4iz

「iPhone 5」、10 月 7 日に発売か-業界筋情報 http://t.co/wvOFtpH

4 分析の手法

本章では、収集したデータの分析方法、およびデマを検出する手法についてべる。4.1 節でツイート中から感情を表す単語を検出する手法、4.2 節でツイートの伝播ネットワークを求める手法、4.3 節でデマとニュースを分類するために用いる指標、および 4.4 節でデマを検出するアルゴリズムについて述べる。

4.1 肯定語・否定語の検出

ツイート中の肯定語の数、否定語の数は、文献 [9] においてデマとニュースの識別のために有用な指標とされている。

本研究では、収集したツイートから自動で肯定語および否定語を検出するシステムを作成した。このシステムにおいて肯定語と否定語の判別には、乾らが公開している日本語評価極性辞書 [14] を利用している。この辞書は、用言約 5200 語および名詞約 8300 語に対し、その単語の持つ意味がポジティブであるか、ネガティブであるかが記載されている [15][16]。

この辞書を利用して、以下に示す手順でツイート中に存在する肯定語、否定語の数を求めた。

1. 形態素解析
   形態素解析とは、文章を単語ごとに分割し、その品詞を求める作業である。本研究では、茶筌 [17] と呼ばれる形態素解析システムを用いて、ツイートを単語ごとに分割した。

2. 評価極性辞書との比較
   ツイート中に含まれる単語から、肯定語および否定語を検出した。

3. 出力
   検出した肯定語、否定語それぞれの数、および検出単語を出力した。

4.2 ツイート伝播経路の取得

本研究では、Twitter が公式にサポートしているリツイートサービスの機能を利用し、ツイートがユーザ間をどのように伝播したかをツリー状で表現した (リツイートツリー)。その際、データ取得には Twitter API を利用し、結果をテキスト形式で取得した。ノードはユーザを、エッジはリツイートの挙動を表現している。

4.3 分類の指標

本研究で収集したツイートのユーザ情報 (リツイートした人のフォロワー数、ツイート数など)、肯定語・否定語数、リツイートツリーを分析した結果、以下に示す 3 つの指標についてデマとニュース間の有意差が見られた (t 検定・有意水準 5 %)。そのため、これらの指標を用いてデマとニュースの分類を行う。

- 否定語数 + 肯定語数 : $\boldsymbol{Np}$
  拡散したデマはニュースと比較してネガティブ単語・ポジティブ単語が多く含まれている傾向が見られた。

- 1 – エッジ数/ノード数 : $\boldsymbol{1 - En}$
  デマのリツイートツリーはニュースと比較して、親ノードを持たず独立するノードが多く見られた (図 1、図 2)。これは、デマのリツイートツリーがニュースと比較してノード数に対するエッジ数が少ないことを表す。

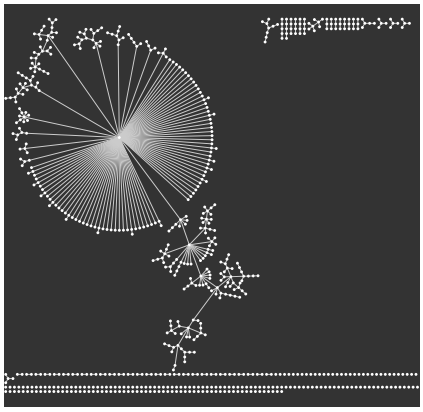

図 1: デマのリツイートツリーの例

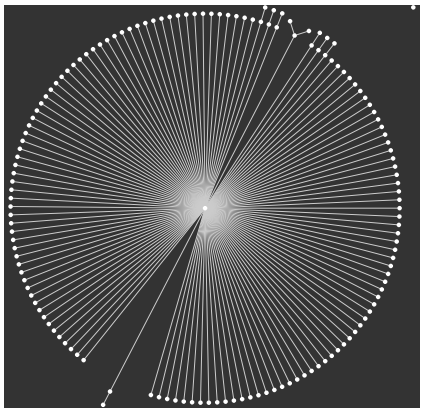

図 2: ニュースのリツイートツリーの例

- リツイート最大深さ/ノード数：***Rn***
  文献 [9] より、リツイート最大深さが信用性に影響を与えると報告されている。今回取得した情報でも、デマほどリツイート最大深さが大きくなる傾向が見られた。今回取得したリツイートツリーは、ツイートごとにノード数が異なるため、リツイート最大深さをノード数で割って正規化している。

図 3〜図 5 に上記 3 つの指標を用いたデマとニュースの比較結果を示す。

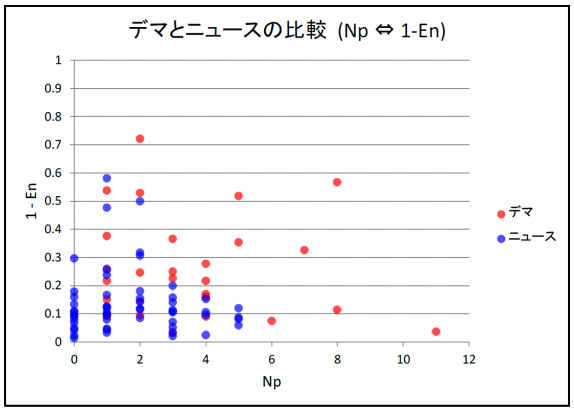

図 3: デマとニュースの比較 ($Np \Leftrightarrow 1 - En$)

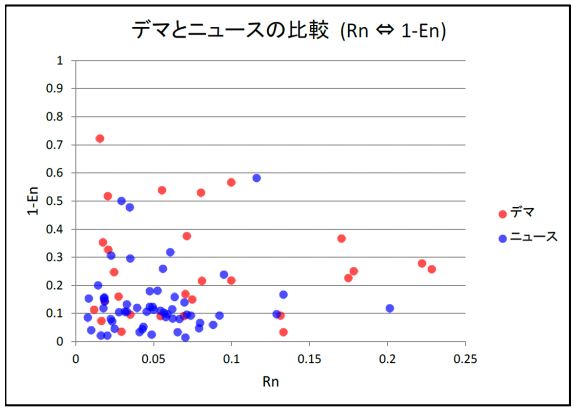

図 4: デマとニュースの比較 ($Rn \Leftrightarrow 1 - En$)

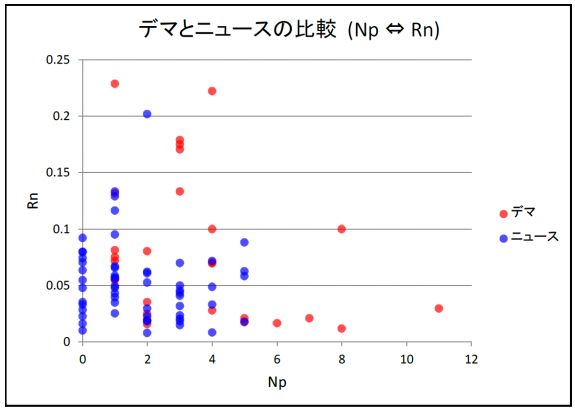

図 5: デマとニュースの比較 ($Np \Leftrightarrow Rn$)

4.4 分類アルゴリズム

上記の 3 つの指標を用いてデータをプロットすると、ニュースクラスタとそれ以外のノイズとみなすことができる。そのためツイートの分類には、ノイズとクラスタに分類可能なノイズクラスタリング [18] を用いた。

5 結果

本章では、与えられたツイートデータに対してデマかどうかの分類を行った結果、およびツイートのデマの度合を算出するする手法とその適用結果について述べる。

5.1　デマの検出

4.3 節で述べた 3 つの指標について、ノイズクラスタリングを実行した。入力データは、0〜1 の範囲で正規化している。

デマ、ニュースの検出率を表 2 に、結果のグラフを図 6、図 7 に示す。

表 2: 検出率

	デマ	ニュース	全体
T	19 (64 %)	48 (89 %)	67 (80 %)
F	11 (36 %)	6 (11 %)	17 (20 %)

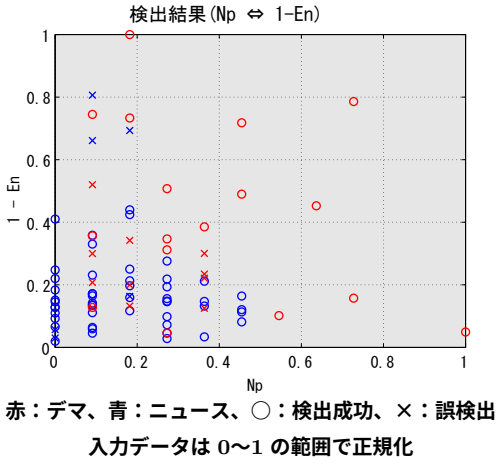

図 6: 検出結果 ($Np \Leftrightarrow 1 - En$)

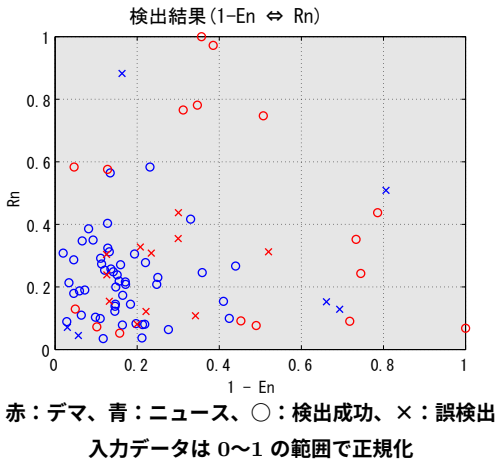

図 7: 検出結果 ($1 - En \Leftrightarrow Rn$)

5.2　デマ度の推定

図 6、図 7 より、ニュースは原点周辺に集中しており、デマは全体に分散している傾向が見られる。そこで、原点からの距離を利用して、ツイートの「デマ度」を推定する関数を構築した。関数は以下の式で表される。

$$D = \sqrt{((\alpha Np)^2 + (\beta(1 - En))^2 + (\gamma Rm)^2)}$$

D：デマ度、α、β、γ：正規化のための係数

$\alpha = 0.0909$、$\beta = 1.38$、$\gamma = 4.38$

この式を収集したデータに当てはめた結果を図 8 に示す。計算結果の平均値は、デマ：$D = 0.69$、ニュース：$D = 0.39$ であった。

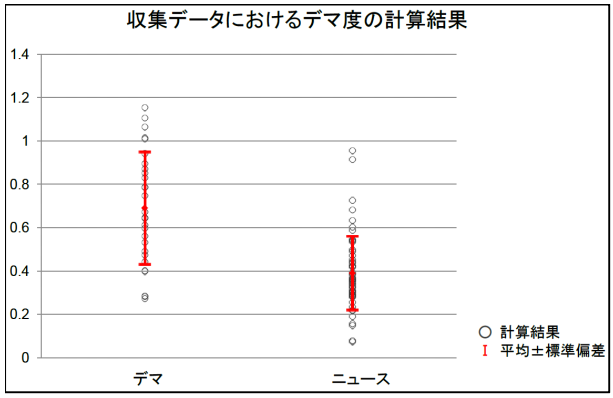

図 8: 収集データにおけるデマ度の計算結果

6　まとめ

我々は Twitter 上を伝播するツイートを収集し、教師なし分類によってデマとニュースの分類を行った。その結果、リツイートの多いツイートはニュースクラスタとその他のノイズとして分類されることがわかり、有意な特徴量を抽出することで、学習コストを抑え高い精度の分類を実現した。これにより、ツイートを精度 80 %で自動的にデマとニュースに分類可能となった。

今後は、非公式なリツイートによって伝播されるツイートの経路についても考慮を行い、より精度を高めていく必要がある。また、ツイートがデマである度合をユーザにフィードバックする仕組みについても検討を行う必要がある。

参考文献

[1] T. Sasaki, M. Okazaki, Y. Matsuo: “Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors”, SOM (2010).

[2] 震災に伴うメディア接触動向に関する調査 (野村総合研究所)
http://www.nri.co.jp/news/2011/110329.html

[3] 藤江利彦: ‘はじめてのマスコミ論’, 同友館 (2006).

[4] 香内三郎, 山本武利: ‘現代メディア論’, 新曜社 (1987).

[5] 佐藤卓己: ‘現代メディア史’, 岩波テキストブックス (1998).

[6] 鈴木みどり: ‘メディア・リテラシーの現在と未来’, 世界思想社 (2001).

[7] 富山英彦: ‘メディア・リテラシーの社会史’, 青弓社 (2005).

[8] M. Schmierbach, A. Oeldorf-Hirsch: “A little bird told me, so I didn ’t believe it: Twitter, credibility, and issue perceptions”, AEJMC (2010).

[9] C. Castillo, M. Mendoza, B. Poblete: “Information Credibility on Twitter”, Proceeding of the 20th international conference on World wide web, pp.675-684 (2011).

[10] J. Ratkiewicz, M. Conover, M. Meiss: “Detecting and Tracking the Spread of Astroturf Memes in Microblog Streams”, arXiv (2010).

[11] 早川洋行: ’ 流言の社会学’, 青弓社 (2002).

[12] 川上善郎, 松田美佐, 佐藤達哉: ‘うわさの謎’, 日本実業出版社 (1997).

[13] Twitter API wiki
http://usy.jp/twitter/index.php?Twitter%20API

[14] 日本語評価極性辞書
http://cl.naist.jp/ inui/research/EM/sentiment-lexicon.html

[15] 小林のぞみ, 乾健太郎, 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, Vol.12, No.2, pp.203-222 (2005).

[16] 東山昌彦, 乾健太郎, 松本裕治: “述語の選択選好性に着目した名詞評価極性の獲得”, 言語処理学会第 14 回年次大会論文集, pp.584-587 (2008).

[17] 茶筅
http://chasen.naist.jp/hiki/ChaSen/

[18] 荒井健太: “逐次的クラスター抽出アルゴリズムの開発と比較検討”, 筑波大学大学院システム情報工学研究科修士論文 (2009).