フェイクニュース分類器を用いた口コミサイトのレビューの分析

岡山　光平† 　石川　博†† 　廣田　雅春†

† 岡山理科大学 総合情報学部 情報科学科　〒 700–0003 岡山県岡山市北区理大町 1–1
†† 首都大学東京 システムデザイン学部 情報通信システムコース　〒 191–0065 東京都日野市旭が丘 6–6
E-mail: †i14i020ok@ous.jp, ††ishikawa-hiroshi@tmu.ac.jp, †††hirota@mis.ous.ac.jp

あらまし　近年，読者を欺く意図を持って作られる「フェイクニュース」が社会問題となっている．フェイクニュースは，SNS などで情報が拡散されやすいため，影響力が大きい．一般的に，ニュースの情報の真偽の判別には，多くの時間コストや，その分野に関する知識が必要であるため，フェイクニュースに含まれる情報の真偽の判定は困難である．そこで，フェイクニュースの情報の真偽を自動的に判別する技術の開発が期待される．本研究では，フェイクニュースとフェイクでないニュースの間には，ニュースのテキストに用いられる表現の特徴に差異があるという考えに基づいて，フェイクニュースを発見することに取り組む．本論文では，Fake News Challenge のデータと政治に関するニュースを SVM により学習することで，フェイクニュース分類器を作成する．フェイクニュースの特徴を機械的に学習できれば，ニュース以外のテキストに対しても，同様に表現の差異があれば，フェイクの内容を含むテキストを検出可能であると考え，レビューを分類する．本論文では，フェイクニュース分類器を Yelp のレビューに適用することで，フェイクの内容を含むと思われるレビューを発見し，その結果を考察する．

キーワード　テキスト分析，文書分類，SVM，Yelp

1. はじめに

Facebook(注1) や，Twitter(注2) などの SNS (Social Networking Service) の普及により，ユーザは多様な情報を取得・共有・発信することが可能になった．これらの SNS では，スマートフォンや，タブレットを用いて，いつでも，どこでも，簡単に情報のやり取りが可能になったため，情報がこれまでよりも速い速度で拡散するようになった．マスメディアも，自社のニュース記事を多くの読者に読んでもらうために，SNS で発信を行なうなどマスメディアでの SNS の活用が広がっている．しかし，これに合わせて，「フェイクニュース」が社会問題となっている．フェイクニュースとは，虚偽の情報でつくられたニュースのことである (注3)．

フェイクニュースの内容は，政治的な内容などの読者の興味を引きやすい扇情的なのものが多く，SNS で拡散されやすい．実際に，2016 年 11 月に行われたアメリカ合衆国大統領選挙からフェイクニュースは爆発的に増加し，社会的な問題となった．フェイクニュースの目的として，サイトの閲覧数を増加させることで，広告収入を増加させることなどがあげられる．フェイクニュースが含む誤った情報が拡散されると，特定の個人や団体が誹謗・中傷の対象となる場合や，不利益をもたらす可能性がある．そのため，フェイクニュースの拡散や，誤った情報を含むコンテンツの閲覧などを防止する必要がある．

一般的に，フェイクニュースなどの情報の真偽の判別には，そのニュースに関する多くの情報の解釈や，多くの時間コストが必要であるため，フェイクニュースの情報の真偽の判定は困難である．そこで，情報の真偽を自動的に判別する技術の開発が期待される．

本研究では，フェイクニュースとフェイクではないニュースの間には，ニュースのテキストに用いられる表現の特徴に差異があるのではないかと考える．近年，Fake News Challenge(注4) などのフェイクニュースを発見するための活動が盛んに行われている．Fake News Challenge とは，多くのニュース記事の中からフェイクを含んだニュースを判別するために，人工知能や，機械学習，自然言語処理などの技術がどのように活用できるかを探求するプロジェクトである．そこで，本研究では，Fake News Challenge のフェイクニュースとフェイクではないニュースを教師あり学習によって学習することで，フェイクニュースのテキストで用いられる表現の特徴を学習し，分類器を作成する．さらに，ニュース以外のテキストでも同様に表現の差異があれば，その分類器によって，フェイクの内容を含むテキストを検出することが可能なことを検証する．

本研究では，ニュース以外のテキストとして，口コミサイト Yelp(注5) に投稿されたレビューを用いる．口コミサイトでは，ある特定の対象に対する意見を記述したレビューが，多数のユーザによって投稿される．そして，そのレビューの内容によって，他のユーザの購買意欲や，レビューの対象に対しての評価に影響を与える．また，レビューの内容が金銭上の利益につながるため，レビューは重要な役割を持っている．しかしながら，その利益を目的としたスパマーによるスパムレビューの投稿が問題になっている．スパマーは，商品や店の評判を意図

(注1)：https://ja-jp.facebook.com/
(注2)：https://twitter.com/
(注3)：https://kotobank.jp/word/フェイクニュース-1748301
(注4)：http://www.fakenewschallenge.org/
(注5)：https://www.yelp.com/

的に上下させることを目的に，悪意を持ってスパムレビューを投稿する．そのため，フェイクな情報を含むレビューもスパムとして考えることが可能である．スパムレビューも，フェイクニュースと同様に，虚偽の情報で作られたテキストであり，また，その真偽の判定も同様に困難である．そこで，本研究では，フェイクニュースの分類器を用いて，レビューをフェイクかどうかを分類し，その結果を分析する．

本論文の構成は以下のとおりである．第 2 節では，フェイクニュースの分析に関連する研究として，誤情報の拡散とレビューの分類の研究について述べる．第 3 節では，フェイクニュース学習器で口コミサイトのレビューの分析を行う手法について述べる．第 4 節では，本研究で提案するフェイクニュース分類器による Yelp のレビューからフェイクレビューの検出の分析結果について述べる．第 5 節では，フェイクニュース分類器によって分類されたフェイクレビューとフェイクでないレビューを比較した結果について述べる．第 6 節では，本論文で得られた成果をまとめ，今後の課題について述べる．

2. 関 連 研 究

2.1 誤情報の拡散防止についての研究

本研究で扱うフェイクニュースに関連するものとして，デマや流言などと呼ばれる誤情報があげられる．誤情報の拡散防止に関する研究として，2011 年に起こった東日本大震災をはじめとした震災時のデマが扱われることが多い．その際に Twitter で拡散されたデマについて，多方面から研究がなされている．

梅島ら [8] は，デマの拡散を防止するため，災害時の Twitter における，デマを含むツイートと，デマの訂正を含むツイートの拡散に関する仮説を立て，それらのツイートの印象をポジティブ，不安を煽るなどの評価属性をもとに評価し，傾向を分析した．中原ら [6] は，リツイートの回数や，コメント付きリツイートに関するコメントを用いることで，ユーザに訂正ツイートとその情報の危険度を提示し，情報がデマであるかの判断を支援する手法を提案した．渡邉ら [7] は，誤情報に多く含まれるキーワードで収集されたツイートを，誤情報の支持・拡散ツイート，反論・訂正ツイートに分類するためのコーパスを構築し，教師あり学習を用いてツイートの内容が誤情報か反論であるかを自動分類するための手法を提案した．

Twitter は，誤情報を拡散するためのツールとして非常に有力なので，デマツイートなどの誤情報の拡散を防止する研究は多くなされている．今後，Twitter 上でのフェイクニュースの拡散などについて，本研究で取り組むフェイクニュース分類器などが応用可能な一分野であると考える．

2.2 レビューの分類についての研究

これまでに，商品や映画に対する評価についてのレビューを用いて，情報の信頼性や，レビューの内容について分類を行う研究は多くなされている．

山澤ら [4] は，Amazon(注6) で公開されているカスタマーレビューを用いて，ユーザが内容を信頼して利用できるレビューを自動抽出することを目的とした判別実験を行った．松本ら [5] は，レビューの文章の単語間の関係についての情報を含む単語の出現パターンをパターンマイニングの手法を用いて抽出し，文書内に現れる単語の出現頻度とともに分類に用いた場合の分類性能への影響について，実験により映画の英文レビューの肯定，否定への分類を調査した．

これらの研究では，ユーザにとって有用なレビューであるか，内容が肯定的か否定的かというところに注目して，それぞれ分類を行っている．本研究では，レビューの内容がフェイクか否かという点に注目して，分類を行う．

本研究では，フェイクニュース分類器を用いて，Yelp に投稿されたレビューをフェイクかどうかを分類する．Yelp には，フェイクであると疑わしいレビューを自動的にフィルタリングする機能があるが，そのアルゴリズムは公開されていない．Arjun ら [2] は，フィルタリングされたレビューとされていないレビューを用いて，Yelp のフェイクレビューのフィルターが，どのような動作をしているかについて仮定を立てて，調査を行った．本研究では，Yelp が公開しているレビューを用いるため，Yelp によって，フェイクレビューのフィルターが適用された状態であると考えられる．そこで，本研究では，フェイクレビュー分類器を適用することで，新たなフェイクレビューや，フェイクレビューと分類されるレビューを分析することで，新たな知見を獲得することに取り組む．

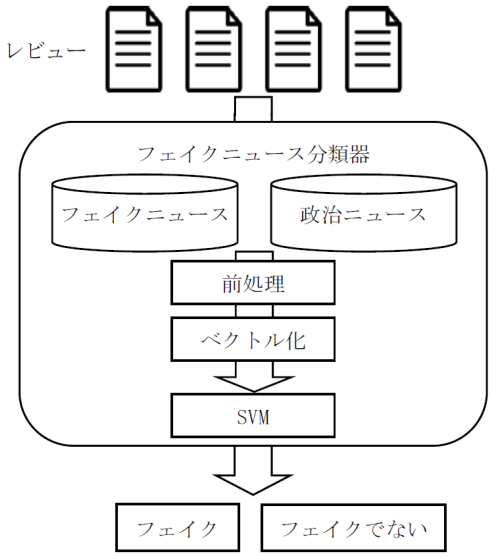

図 1　提案手法の概要

3. 提 案 手 法

本節では，フェイクニュースを学習した分類器で口コミサイトのレビューの分析を行う手法について述べる．図 1 に，提案手法の概要を示す．提案手法では，まず，フェイクニュースとフェイクではないニュースの 2 種類のデータをトレーニングデータとすることで，フェイクについての学習を行い，フェイクニュース分類器を作成する．その後，Yelp に投稿されたレビューに対してフェイクニュース分類器を適用することで，そ

(注6)：https://www.amazon.co.jp/

のレビューがフェイクかフェイクでないかの分類を行う．そして，フェイクと判別されたレビューの分析を行う．

3.1 前 処 理

本節では，トレーニングデータの作成と，前処理について述べる．フェイクニュースのデータとフェイクではないニュースのデータをトレーニングデータとする．ここで，本論文で扱うトレーニングデータに含まれるテキストのすべては，英語で記述されたものである．テキストの言語判定には，langdetect(注7) を用いた．その後，それらのテキストに対して，前処理として，各単語の小文字化，ストップワードの除去，ステミングを行う．ストップワードの除去には，多くのテキストで使われており一般的な「the」,「and」などの単語を 681 語指定したストップワードリストを用いる．

3.2 ベクトル化

本節では，トレーニングデータのテキストをベクトル化する手法について述べる．はじめに，3.1 節で前処理を行ったトレーニングデータのテキストに対して，TF-IDF を適用し，ベクトル化を行う．そして，TF-IDF の次元数を削減するために，PCA (Principal Component Analysis) [3] を適用する．また，PCA で得られた特徴量に対して標準化を行い，その結果を以下の処理では用いる．

3.3 フェイクレビューの判別

本節では，フェイクニュース分類器の作成と，Yelp のレビューからフェイクを含むレビューを判別する手法について述べる．本研究では，教師あり学習の手法の SVM (Support Vector Machine) [1] を用いて，フェイクニュース分類器を作成する．その際に，SVM では，レビューをフェイクか，フェイクでないかの 2 値分類を行う．はじめに，3.2 節で作成したベクトルを SVM で学習し，フェイクニュース分類器を作成する．

次に，得られたフェイクニュース分類器を適用することで，口コミサイトのレビューがフェイクかを判別する．口コミサイトのレビューに対しても，同じ手順でベクトルを作成し，その結果をフェイクニュース分類器に適用する．そして，フェイクと分類されたレビューについて分析する．

4. フェイクニュース分類器の作成

本節では，フェイクニュースとフェイクではないニュースから生成したフェイクニュース分類器の性能の評価を行う．

4.1 実験で用いる SVM について

本実験では，SVM を用いて，フェイクニュースとフェイクでないニュースについて学習し，フェイクニュース分類器を作成する．そのため，SVM のパラメータのチューニングと，PCA の適切な次元数を決定する必要がある．SVM のハイパーパラメータである，c と γ と，PCA の次元数については，5 交差検定によるグリッドサーチを用い，F 値が最も高いものを用いた．また，本実験では，SVM のカネール関数にガウスカーネルを用いた．

(注7)：https://pypi.python.org/pypi/langdetect

表 1 フェイクニュース分類器の実験結果

	Precision	Recall	F 値	Accuracy
フェイク	1.000	1.000	1.000	0.996
フェイクでない	0.993	1.000	0.996	
平均	1.000	1.000	1.000	

4.2 データセット

本節では，実験に使用するデータセットについて述べる．

本研究では，分類器の学習に使用するフェイクニュースのトレーニングデータとして，フェイクであると人手で判断されたニュースとフェイクではない米国の政治ニュースの 2 種類のデータを用いた．

フェイクのニュースのデータには，Fake News Challenge で公開されたデータセットの中でフェイクであるとされているニュースのデータセットを用いる．Fake News Challenge に含まれるフェイクニュースの件数は，1,679 件である．

フェイクではない政治のニュースのデータには，webhose.io(注8) で公開されている米国の政治を扱ったニュースのデータセットを用いた．前述したように，フェイクニュースの内容の多くは政治的なニュースであるので，フェイクではないニュースにも政治に関するニュースデータセットを用いた．

この webhose.io のデータセットは，87,157 件のニュースを含んでいるが，実験に適切でないテキストを含むニュースを取り除く．webhose.io の中でニュースの本文が存在しない，また，本文の長さが極端に短いものを除去するために，本文の単語数が 50 語未満，または本文の一部が省略されているニュース (文末が「...」であるニュース) を取り除いた．この処理の結果，フェイクではない政治に関するニュースの件数は， 24,088 件である．これらのデータセットを用いてフェイクニュース分類器を作成する．

4.3 実 験 結 果

表 1 に，フェイクニュース分類器の Precision，Recall，F 値，および Accuracy の実験結果を示す．この表では，フェイクは Fake News Challenge のフェイクニュースを，フェイクでないは webhose.io のフェイクでない政治ニュースを示す．表 1 の結果より，フェイクニュース分類器の性能はいずれの評価指標においても高い値になっている．フェイクニュース分類器が高い性能を示しているが，過学習の可能性もあるので，今後検討する．以下では，このフェイクニュース分類器を用いて，レビューがフェイクを含むかどうかを分析する．

5. Yelp のレビューの分析

4. 節で作成したフェイクニュース分類器を用いて口コミサイトのレビューを分析する．

5.1 データセット

本論文では， フェイクニュース分類器を適用するテストデータとして，Yelp Dataset Challenge (round 9)(注9) を用いる．このデータには，144,072 件の Venue と，4,153,150 件のレビュー

(注8)：https://webhose.io/

(注9)：https://www.yelp.com/dataset/challenge

表 2　Yelp のレビューの分類結果

	件数
フェイク	3,148
フェイクでない	996,852
全件	1,000,000

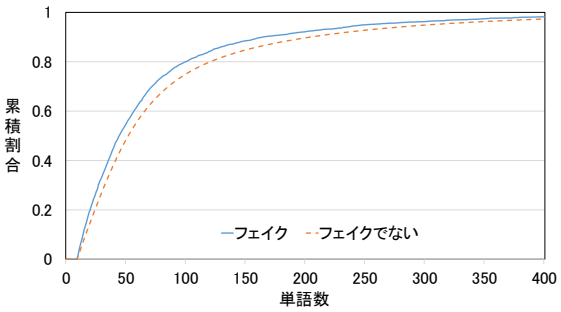

図 2　テキストの長さの累積分布

が含まれている．このデータのレビューの中から，トレーニングデータと同様に英語で記述されたレビューを，ランダムに 1,000,000 件を選択し，テストデータとする．

テストデータのレビューについても，トレーニングデータと同様に，前処理とベクトル化を行う．その後，フェイクニュース分類器を適用することで，レビューがフェイクかどうかを分類する．

5.2　実験結果

表 2 に，フェイクニュース分類器による Yelp のレビューの分類結果を示す．この結果では，フェイクと判別されたレビューが 3,148 件で，約 0.3%ある．Yelp では，多くのレビューが投稿されるが，Yelp 自体がスパムなどの悪意のあるレビューを排除することや，口コミサイトなので一般のレビューがレビューの大半を占めるため，フェイクと判別されるレビューが少なかったと考えられる．

5.3　フェイクレビューとフェイクでないレビューの比較

この節では，フェイクニュース分類器によって，フェイクと分類されたレビューと，フェイクでないと分類されたレビューについて，言語的な特徴を分析し，比較する．

5.3.1　テキストの長さの比較

本節では，レビューのテキストの長さに着目し，フェイクと分類されたレビューと，フェイクでないと分類されたレビューの比較を行う．ここでのテキストの長さとは，1 つのレビューの単語の総数とする．フェイクと判別されたレビューとフェイクでないと判別された全てのレビューのテキストから，1 レビューあたりの平均の単語数を算出した．

結果として，フェイクと判別されたレビューの単語数の平均値は，約 76.1 語であった．また，フェイクでないと判別されたレビューの単語数の平均値は，約 89.2 語であった．

フェイクと判別されたレビューと，フェイクでないと判別されたレビューのレビューの単語数の累積分布を図 2 に示す．図 2 において，フェイクでないと判別されたレビューの方が単語数が短いことがわかる．フェイクと判別されたレビューの方が

表 3　出現確率に偏りのある単語 上位 10 語

	フェイク	フェイクでない
1 位	roll	pizza
2 位	taco	experi
3 位	star	wait
4 位	sushi	store
5 位	pretti	friend
6 位	server	sandwich
7 位	menu	recommend
8 位	restaur	beauti
9 位	chicken	move
10 位	beer	spot

平均の単語数が少なく，レビューのテキストが短い傾向である．これは，フェイクな情報を含んだスパムレビューを書くスパマーが，短い時間で大量のスパムレビューを書くため，レビュー 1 件あたりのテキストが短くなるためだと考えられる．

5.3.2　単語の出現頻度の比較

本節では，レビューに用いられている単語の出現頻度に着目し，フェイクと分類されたレビューと，フェイクでないと分類されたレビューの比較を行う．それぞれのレビューから，単語の出現確率を算出し，その差分を求めることで，フェイクとフェイクでないレビューのどちらかに偏って出現する単語を求めた．

表 3 に，フェイクと分類されたレビューと，フェイクでないと分類されたレビューのどちからに出現確率が偏っている単語のそれぞれの上位 10 件を示す．この結果の中の単語は，3.1 節の前処理を適用後なので，ステミングやストップワードの処理が行われた後の状態である．

表 3 において，フェイクと判別されたレビューでは，‘ recommend ’ , ‘ friend ’ といった単語の出現頻度が低いことが分かった．また，‘ fake ’ という単語は，フェイクと判別されたレビューにおいて，出現頻度が低い．実際に，‘ fake ’ という単語を用いたレビューは，テストデータ全体では約 0.26%であったのに対し，フェイクと判別されたレビューでは約 0.16%であった．これは，ユーザが意図的に嘘のレビューを書く際に，閲覧者に嘘であることを思い浮かばせないように，レビューのテキストに ‘ fake ’ という単語を用いることを意図的に避けている可能性がある．

5.3.3　URL を含むレビュー数の比較

本節では，テキスト中に URL を含むレビューに着目し，フェイクと分類されたレビューと，フェイクでないと分類されたレビューの比較を行う．

梅島ら [8] は，デマのツイートに関して，URL を含む RT はデマである可能性が低いことを示した．これは，情報源のサイトの URL を含んでいるツイートは，情報の裏付けがあり信頼できるため，デマである可能性が低いためである．このことから，デマのテキストの判別には，テキストに URL を含むかどうかが有効であると考えられる．そこで，本研究では，レビューのテキスト中に URL を示す ‘ http ’ が含まれているレビューの件数をカウントした．

フェイクと判別されたレビューでは，3,148 件中 5 件が該当

し，割合は約 0.16%であった．また，フェイクでないと判別されたレビューでは，996,852 件中 2,546 件が該当し，割合は約 0.23%であった．結果として，どちらにも URL を含んだレビューは少ないことが分かった．これは，Twitter では投稿するツイートに 140 文字の字数制限があるのに対して，Yelp では投稿するレビューの字数制限が，5,000 文字であるためであると考えられる．

6. おわりに

本研究では，フェイクニュースの情報を学習したフェイクニュース分類器を作成し，Yelp のレビューの分析に用いた．フェイクニュースと政治のニュースからフェイクニュース分類器を作成し，評価実験を行った．また，Yelp のレビューに対してフェイクニュース分類器を適用し，レビューがフェイクかどうかの分類を行った．そして，フェイクと分類されたレビューとフェイクでないと分類されたレビューを比較することで，フェイクなテキストを含むレビューの言語的特徴を分析した．分析として，テキストの長さの比較，単語の出現頻度の比較，URL を含むかどうかについて比較した．

今後の課題として，フェイクニュース分類器の過学習の検討があげられる．また，Yelp のレビュー以外のテキストに対して，フェイクニュース分類器の適用を行うこともあげられる．本研究では，フェイクと分類されたレビューとフェイクでないと分類されたレビューを，テキストの長さの比較，単語の出現頻度の比較，URL を含むかどうかという観点で比較した．しかし，別の観点から比較を行うことで，フェイクなテキストを含むレビューの新たな言語的特徴を分析することが可能であると考えられる．

謝　辞

本研究は，首都大学東京傾斜的研究 (全学分) 学長裁量枠戦略的研究プロジェクト戦略的研究支援枠「ソーシャルビッグデータの分析・応用のための学術基盤の研究」，および JSPS 科研費 16K00157, 16K16158 による．

文　献

[1] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, Vol. 20, No. 3, pp. 273–297, 1995.
[2] Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie S Glance. What yelp fake review filter might be doing? In *Proceedings of the 7th International Conference on Weblogs and Social Media*, 2013.
[3] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists*, Vol. 2, No. 1, pp. 37 – 52, 1987.
[4] 山澤美由起, 吉村宏樹, 増市博. Amazon レビュー文の有用性判別実験. 情報処理学会研究報告自然言語処理 (NL), Vol. 2006, No. 53 (2006-NL-173), pp. 15–20, 2006.
[5] 松本翔太郎, 高村大也, 奥村学. 単語の系列及び依存木を用いた評価文書の自動分類. 第 3 回情報科学技術フォーラム (FIT 2004) 講演論文集第 2 分冊 F-006, pp. 213–214, 2004.
[6] 中原英美, 冨永一成, 牛尼剛聡. リツイート構造を用いたデマ拡散防止支援手法. 第 4 回データ工学と情報マネジメントに関するフォーラム, F2-3, 2012.
[7] 渡邉研斗, 鍋島啓太, 岡崎直観, 乾健太郎. Twitter 上での誤情報と訂正情報の自動分類. 言語処理学会第 19 回年次大会, 2013.
[8] 梅島彩奈, 宮部真衣, 荒牧英治, 灘本明代ほか. 災害時 Twitter におけるデマとデマ訂正 RT の傾向. 研究報告 情報基礎とアクセス技術, Vol. 2011, No. 4, pp. 1–6, 2011.