研究指導 中澤 真 教授

流言検出システムの精度向上

ー訂正情報の特徴量に着目してー

飛木 優里

1. はじめに

急速なインターネットの普及に伴い，SNSの利用が活発になっている．2020年度のSNS利用動向に関する調査によると，日本国内のネットユーザの約8割がSNSを利用している[1]．そうしたSNSの1つであるTwitter[1]は，不特定多数の人々と気軽につながることができることや，情報共有や拡散の容易さから多くの人々に利用されている．こうしたSNSの特性は人々の交流を促進する一方で，真偽不明な流言の発信や拡散という諸刃の剣となっている．この問題に対応するため，Twitter上の流言を検出しユーザに伝えるサービスやシステムとして，FIJが提供している物事の真偽を調査・検証している記事を紹介するファクトチェック・ナビ[2]といったものが無償で提供されている．

しかし，これらのサービスは流言を認知してから検証までに時間を要するため，流言か否かの結果をすぐに提示することが難しい．このため，検証を行っている間にも根拠がない情報がSNS上で拡散してしまう恐れがある．

この問題に対して，Twitter上でユーザに流言の可能性があるツイートを提示することで，流言の拡散を防止するような研究が行われている[2][3][4][5]．中でも宮部ら[6]の研究は，流言を指摘するツイートのことである訂正情報を利用し，ツイートを流言か否か判別している．この研究は訂正情報によって流言内容を特定するため，流言のみのツイートに着目するよりも検出精度が向上している．ツイートが訂正情報か否かを判別する特徴量としては，ツイート内の形態素数や，リツイート(以下，RTと呼ぶ)・URLの有無などが用いられている．しかし，訂正情報を判別する条件が限られていることから改善すべき点も多い．判別精度の向上には，RT・URLの有無などのツイートにおける表面的な特徴量だけではなく，ツイートの内容にまで踏み込んだものも追加することが必要である．

そこで本研究では，流言の検出力に直結する訂正情報の判別精度向上を目指し，新たな特徴量を模索する．さらに，この新しい特徴量が先行研究の特徴量よりも高い判別精度を達成できることを明らかにする．

2. Twitterにおける流言検出システム

現在，SNSの利用者は年々増加傾向にあり，特にTwitterは情報交流が容易なことから利用者が多い[1]．しかしこの便利さが流言の拡散にもつながってしまい，情報共有に悪影響を及ぼしている．これに伴い，Twitter上の流言を検出しユーザに伝えるサービスやシステムが必要になり，既存のサービスでは流言情報クラウド[3]というサイトが提供されている．これらはユーザに流言だと気づきを与えることで流言の拡散防止に役立つものであり，最近ではこのようなサービスが，新型コロナウイルスに関する根拠が怪しい情報の真偽を確認するためにも利用されている．

前述した流言情報クラウドでは，訂正情報に着目することで流言を検出している．訂正情報とはTwitterで発信される流言を指摘しているツイートのことであり，宮部ら[6]の研究では，流言の不確かさに言及しているテキストのことと定義されている．訂正情報と判定されるツイートが一定数以上ある事象を流言と判別するのが基本コンセプトであり，このため訂正情報であるか否かの判別精度が，流言検出の性能を左右することになる．

宮部らの手法では訂正情報判別のための特徴量として，ツイートの形態素数やRT・URLの有無を用いている．しかし，これらの特徴量はツイートの表面的な要素のみを扱っており，ツイート内容に関する情報ではない．そのため，訂正情報に出現する特徴的な語に関する情報も特徴量に加えれば，判別精度をさらに改善できる可能性がある．

以上のことを踏まえて，本研究では訂正情報に出現する語の傾向に着目し，訂正情報の判別精度向上の可能性を探る．

3. 訂正情報判別に用いる特徴量

まず，訂正情報判別に適した特徴量を検討するために，流言に関連するツイートを収集し，自然言語処理の技法を用いてこれを分析する．

3.1 ツイートの収集と訂正情報のラベリング

流言に関連するツイートを収集するために，Twitter上で新型コロナウイルスに関する以下の7種類の流言を対象と定めた．これらの流言を信じるツイートもそれを訂正するツイートも分析に必要な件数

[1] https://twitter.com/

[2] https://navi.fij.info/

[3] http://mednlp.jp/~miyabe/rumorCloud/rumorlist.cgi

を十分得られることが選択の理由である．収集対象となるツイートは，取り上げた流言に関する単語の他に，ツイート群を収集する際に用いる，流言を示す表現である流言マーカとして選んだ「デマ，嘘，ウソ，騙され」の4語のいずれかを含んでいるものとして，流言の種類ごとに約300件ずつGoogle Chromeの拡張機能である「ついすぽ[4]」を用いて収集した．

- 新型コロナウイルスのワクチンを打つと5Gに接続される
- お湯を飲むとコロナウイルスが死滅する
- コロナワクチンを打つとBluetoothにつながる
- コロナワクチンを打つときにマイクロチップが埋め込まれる
- コロナワクチンを打った箇所に磁石がつく
- コロナワクチンを打つと遺伝子操作される
- 通天閣にコロナワクチンに関するライトアップがされる

次に，収集したツイートに対して，訂正情報か否かを人手によってラベル付けをした．本研究では，流言の事象が嘘であると理解しているツイートを訂正情報と定義する．ここでは，面白がっているものや揶揄しているものも嘘であると理解しているものとして扱い，対象の流言以外の流言に関しての記述については判断の基準に含めないこととする．また，流言を引用RTした場合も，その事象を流言だと理解しているか否かで判断する．なお，対象の事象に対して疑問表現がある場合は訂正情報に含めないものとする．

3.2 各ツイートの分析と特徴量の決定

定義によりラベル付けした訂正情報1,440件，非訂正情報502件を形態素解析[7]し，それぞれのツイート群についての語の出現頻度を分析した．このテキストデータの統計的分析には，フリーソフトの「KH Coder[5]」とその内部にある形態素解析器である「MeCab[6]」を用いた．2つのツイート群で出現傾向に差異がある語であれば，2つの群を判別するための特徴語となる可能性がある．2つの群での各語の出現傾向の差異を測る評価尺度としては差異係数[8]を用いる．

今回は，訂正情報・非訂正情報の文書数が大きく異なるため，各群の文書数で正規化した値を用いて以下の式のように差異係数を算出した．なお，A群は訂正情報群，B群は非訂正情報群に相当する．

$$差異係数 = \frac{\left(\frac{A群での対象単語の出現頻度}{A群の文書数}\right) - \left(\frac{B群での対象単語の出現頻度}{B群の文書数}\right)}{\left(\frac{A群での対象単語の出現頻度}{A群の文書数}\right) + \left(\frac{B群での対象単語の出現頻度}{B群の文書数}\right)}$$

訂正情報の特徴量となる語（以下，訂正特徴語と呼ぶ）の抽出条件を，差異係数の値が0.50以上かつ訂正情報群での出現頻度が20以上の語であり，特定の流言の内容に依存しない語とした．また非訂正情報の特徴量となる語は，差異係数が－0.50以下かつ非訂正情報群の出現頻度が10以上の語であり，特定の流言の内容に依存しない語と定めた[7]．

これらの条件に基づいた結果を表 1に示す．本研究では，このように抽出した訂正特徴語と非訂正特徴語それぞれの異なり語数[9]を特徴量とし，訂正情報の判別に用いる．

表 1 差異係数により抽出された特徴語

訂正特徴語		
語	出現頻度	差異係数
面白い	30	1.00
便利	29	0.82
本気	21	0.76
www	27	0.65
アホ	26	0.64
悪い	25	0.63
流れる	74	0.62
バカ	24	0.61
流す	127	0.55

非訂正特徴語		
語	出現頻度	差異係数
本当	17	-0.88
もしかして	12	-0.79
現実	11	-0.64
ない	234	-0.64
教える	10	-0.61

ツイート内容の語以外の特徴量としては，リプライの有無・いいねの有無・引用RTの有無を用いる．流言のような真偽が不明な情報は，正しい情報を早急に求める傾向が強まるため，ツイートへの反応率も高くなると考えられる．そのため訂正情報を発信するユーザに，直接真偽を確認したい人がリプライ機能を使用することで，リプライ数が増加すると推測できる．また，流言か否かを認識していない非訂正情報は，同様に流言か否かを疑問視している人の共感を呼び，ツイートに対するいいねや引用RT等の機能で拡散されると予想される．ここで先行研究のRTではなく引用RTを特徴量とした理由は，他人のツイートをそのまま拡散するのではなく，流言に対する自分の意見を添えてツイートする人が多くなるのではないかと推測したためである．

[4] https://tilde.afonomics.com/TweetExport/
[5] https://khcoder.net/index.html
[6] http://taku910.github.io/mecab/
[7] 2 つのツイート群でツイート数に偏りがあるため，しきい値も異なるものとしている．

そのほかに，ハッシュタグの有無も特徴量として検討したが，訂正情報群と非訂正情報群それぞれのハッシュタグの有無の割合に差異が見られなかったため採用を見送った．

4. 先行研究と比較した特徴量の検証

本節では，訂正情報と非訂正情報の判別について，特徴量ごとの判別精度に関して検証する．

4.1 検証前の実施事項

訂正情報群と非訂正情報群のサンプル数の差異が大きく異なるため，判別精度を適切に評価するためには同程度のサンプル数で分析する必要がある．そこで今回は，データ数が少ない非訂正情報群に訂正情報群の件数をそろえ，それぞれ502件で分析する．なお，信頼性の高い結果となるように，K=3としたK分割交差検証法[10]を変則的に用いて訂正情報群を3つに分割し，3通りの分割群の組み合わせで判別分析[11]をした．精度評価はこの平均値を用いる．

ただし，訂正特徴語と非訂正特徴語の異なり語数と，先行研究の特徴量である形態素数は，数量データであるが，そのほかの特徴量はカテゴリデータであるため，ダミー変数[12]として処理して判別分析した．

4.2 先行研究の特徴量の検証

先行研究の特徴量と自分の特徴量の有効性を比較するため，目的変数を訂正情報か否かのラベル，説明変数を先行研究の以下の3つの特徴量として，判別分析による精度評価を実施した．

【先行研究の特徴量】
- 形態素数
- RTの有無
- URLの有無

表 2は先行研究の特徴量を用いた場合の判別的中率，再現率，適合率の分析結果である[9]．再現率の平均値は72.84％と高いが，判別的中率の平均値は61.69％となり，精度には課題があることが確認できた．

表 2 先行研究の特徴量の判別精度

組み合わせ	判別的中率	再現率	適合率
分割群1	62.45%	73.90%	60.13%
分割群2	61.06%	71.12%	59.20%
分割群3	61.55%	73.51%	59.32%
平均値	61.69%	72.84%	59.55%

4.3 自分の特徴量の検証

次に，自分が新たに提案した特徴量の判別精度についても分割交差検証法によって分析した．目的変数は先ほどと同様に訂正情報か否かのラベルとし，説明変数は，先に定めた以下の5つの特徴量とする．

【自分の特徴量】
- リプライの有無
- いいねの有無
- 引用RTの有無
- 訂正特徴語の異なり語数
- 非訂正特徴語の異なり語数

表 3に示した分析結果から，再現率は先行研究よりも低いが適合率が64.96％と高いことで，平均判別的中率が64.87％と先行研究の特徴量を用いた場合よりも，高い精度での判別ができることを確認できた．

表 3 新しく追加した特徴量の判別精度

組み合わせ	判別的中率	再現率	適合率
分割群1	66.73%	64.54%	67.50%
分割群2	64.54%	62.15%	65.27%
分割群3	63.35%	68.53%	62.09%
平均値	64.87%	65.07%	64.96%

4.4 先行研究と自分の特徴量を統合した場合の検証

本節では先行研究と自分の特徴量を統合して訂正情報を判別した場合の精度について検証する．これまでと同様に分割交差検証法による複数回の判別分析を実施し，目的変数はツイートの訂正か否かのラベル，説明変数は8種類すべての特徴量を用いる．

表 4に示したように，先行研究と自分の特徴量を統合した場合の平均判別的中率は，検証した中でもっとも高い精度である66.73％となることを確認できた．また，適合率が先行研究のみの場合と比べて改善されており，訂正情報でないものを訂正情報と判別してしまうことが少なくできることが明らかとなった．また，新しく追加した特徴量のみを用いた場合は，再現率が悪くなってしまっていたが，統合した場合にはこの問題点も改善されていることが示された．

表 4 先行研究の特徴量と新特徴量を統合した場合の判別精度

組み合わせ	判別的中率	再現率	適合率
分割群1	68.82%	72.31%	67.60%
分割群2	65.84%	70.12%	64.59%
分割群3	65.54%	73.51%	63.40%
平均値	66.73%	71.98%	65.20%

5. 標準化判別係数による考察

訂正情報の判別における特徴量は，先行研究の特徴量と新特徴量をすべて統合したものが効果的であるという結果になった．そこで，8種類の特徴量それぞれの判別への影響について検討する．ここでは標準化判別係数に着目して，影響の度合いやそれぞれの特徴量の役割について明らかにする．その結果を表 5に示す．

表 5 各特徴量の標準化判別係数

各特徴量	標準化判別係数
リプライの有無	-0.0188
いいねの有無	0.0100
引用RTの有無	0.3699
RTの有無	0.1905
URLの有無	0.4098
形態素数	-0.1169
訂正特徴語の異なり語数	-0.6315
非訂正特徴語の異なり語数	0.3794

この結果の標準化判別係数における絶対値から読み取れるように，引用RTの有無，URLの有無，各特徴語の異なり語数が訂正情報へ大きく影響し重要であることがわかる．また，訂正情報群の重心は負の値，非訂正情報群の重心は正の値となっているため，標準化判別係数が相対的に大きな負の値となっている訂正特徴語の異なり語数は，語数が多いほど訂正情報と判別される可能性が高くなることが示されている．一方，相対的に大きな正の値の係数に着目すると，引用RTの有無，URLの有無，非訂正特徴語の異なり語数が含まれており，それぞれの有無の割合および非訂正特徴語の異なり語数が多いツイートは，非訂正情報と判別される可能性が高くなることが同じく示されている．

以上のことから，先行研究と自分の特徴量を合わせた特徴量が，訂正情報を判別する上でもっとも有効であると結論づけられる．

6. むすび

本研究では，Twitterにおける新型コロナウイルス関連の，7種類の流言に関するツイートの収集を行い，訂正情報群と非訂正情報群それぞれのツイートに含まれる語に着目して特徴量を構成した．また，それらの特徴量を用いた場合の訂正情報か否かの判別精度について検証した．その結果，語の出現頻度を考慮していない従来の特徴量よりも，本研究の特徴量を用いた場合のほうが高い精度を実現できることを確認できた．また，先行研究の特徴量と自分の特徴量を組み合わせて用いることで，さらに精度が向上することも判明した．このように，訂正情報か否かの判別精度が向上することで流言検出力も高まり，最終的な流言拡散防止につながると予想される．

課題としては，非訂正情報を訂正情報と誤判別してしまう適合率の改善が挙げられる．誤判別の原因としては，非訂正情報が流言を発信しているものから疑問視しているものまでと表現が多様であることや，今回の分析において訂正情報と比較して非訂正特徴語の異なり語数が十分に得られなかったことなどが考えられる．よって，今後は分析対象のデータ数を増やすことで，様々な種類の非訂正情報に共通して出現する語の傾向を明らかにし，判別精度の向上を目指す．

参考文献

[1] 2020 年度 SNS 利用動向に関する調査，ICT 総研，https://ictr.co.jp/report/20200729.html/ （参照 2022-02-07）
[2] 鳥海不二夫ほか，”ソーシャルメディアを用いたデマ判定システムの判定制度評価”，デジタルプラクティス，3 巻 3 号 pp-201-208，2012.
[3] 西村涼太ほか，”情報の信頼性の関心を高める流言注意喚起ボットの開発”，ワークショップ 2020，pp-43-50，2020.
[4] 梅本美月ほか，”Web ページに含まれる流言情報への気づきを与える提示方法の検討”，情報処理学会論文誌，62 巻 1 号 pp-183-192，2021.
[5] 池田圭祐ほか，”口コミに着目した情報拡散モデルの提案及びデマ情報拡散抑制手法の検証”，情報処理学会論文誌数理モデル化と応用（TOM），11 巻 1 号，pp-21-36，2018.
[6] 宮部真衣ほか，”人間による訂正情報に着目した流言防止拡散サービスの構築”，情報処理学会論文誌，55 巻 1 号，pp-563-573，2014.
[7] 末吉美喜，テキストマイニング入門：Excel と KH Coder で分かるデータ分析，オーム社，2019.
[8] 小林雄一郎，ことばのデータサイエンス，朝倉書店，2020.
[9] 金明哲，テキストアナリティクスの基礎と実践，岩波書店，2021.
[10] 高木章光，鈴木英太，最新データサイエンスがよ～くわかる本，秀和システム，2019.
[11] 菅民郎，例題と Excel 演習で学ぶ多変量解析―回帰分析・判別分析・コンジョイント分析編―，オーム社，2016.
[12] 菅民郎，Excel で学ぶ多変量解析入門，オーム社，2013.