The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

解釈性のあるフェイクニュース検出器の実装と評価

Implementation and Evaluation of an Interpretable Fake News Detector

山本 和矢 *1　小山 聡 *2　栗原 正仁 *2
Kazuya Yamamoto*1　Satoshi Oyama　Masahito Kurihara

*1北海道大学工学部
School of Engineering, Hokkaido University

*2北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

Interpretability is an important element of fake news detection so that readers can assess the credibility of news by themselves. We implemented a naive Bayes fake news detection model proposed by Granik and Mesyur and evaluated it with the LIAR dataset in terms of recall, effect of stop words, and interpretability. The recall was affected by the imbalanced data and eliminating stop words did not improve the accuracy but slightly deteriorated it. Some high probability words were interpretable as reasons for fake news but longer phrases had better be considered as clues for fake news.

1. はじめに

SNS の普及により人々が手軽に情報を発信したり，入手することが可能になり我々の生活はより利便性を増した．一方でそうした情報の中には信憑性に欠けたものが含まれている．虚偽の情報でつくられたニュースをフェイクニュースと呼ぶ．フェイクニュースの目的は人を欺くことや衝撃的なタイトルで閲覧数を稼ぎ広告収入を得ること，また特定の人物や団体の根拠のない誹謗中傷をすることである．近年，このフェイクニュースを発端として様々な問題が発生している．その例として，2016 年 12 月 4 日にアメリカ大統領選中に SNS で拡散されたクリントン候補の陰謀説に関するフェイクニュースの真相を確かめようと銃を持った男がレストランに侵入するという事件が起こった *1．

あるニュースがフェイクニュースであるか人手で判別するには内容が虚偽であるか精査しなければならないため時間がかかる．また，メディアの多様化により大量のニュースが生み出されている．それゆえフェイクニュースをすべて人手で検出することは非常に困難である．そうしたことから，フェイクニュース自動検出の研究に対するニーズが高まっている．

一方で，フェイクニュースの自動検出器が構築できたとしても，それを無条件に信頼することにも問題がある．フェイクニュースの検出に誤りがあったり，検出器自体が「フェイク」である可能性も完全には否定できないからである．必要に応じて，読者が検出のプロセスを精査できることや，読者自身がニュースを批判的に読む能力を身に付けることも重要である．そのためには，なぜあるニュースがフェイクだと判定できたのか，その原因を調べることが可能な，解釈性の高い検出器が必要となる．

フェイクニュースの自動検出の試みの一つとして，[Granik 17] はフェイクニュースがスパムメッセージと類似点があるという仮定の元で, スパムフィルターに用いられるナイーブベイズ分類器を用いてフェイクニュースの自動検出を行い 74 %の精度でフェイクニュースを分類している. しかし，これには「データの偏ったデータでしか実験していない」,「ストップワードの除去を行っていない」,「使用しているモデルの解釈性が高いにもかかわらず検出の要因を分析していない」といった課題が存在している．

本研究の目的は [Granik 17] の用いた手法を実装し，ラベルの偏りによりどのような違いが生じるか調べること，ストップワードの除去を行い分類結果の変化を検証すること，フェイクニュース検出の要因となった単語を分析することである．

2. 関連研究

[Granik 17] はフェイクニュースがスパムメッセージと，（1）文法的な間違い，（2）感情的に誇張された表現，（3）読者の意見を操作，（4）内容が虚偽，（5）限定的な類似単語の使用，のような類似点があるという仮定の元で, スパムフィルターに用いられるナイーブベイズ分類器を用いてフェイクニュースの自動検出を行った．

この研究において使われたデータセットはラベルに偏りがありフェイクニュースは全体の 5 %ほどしか含まれていない．[Granik 17] はラベルの偏りが原因で検出の再現率が低くなったと主張している．また，ストップワードの除去を行えば分類の精度が向上すると主張している．しかし，いずれの主張も検証されておらず仮説にすぎない．また，単語的特徴から解釈性の高いモデルを用いてフェイクニュースの分類を行っているにもかかわらず，分類の要因となった単語を分析していない．

その他にもフェイクニュースの自動検出の研究が盛んに行われている．フェイクニュースの自動検出の手法として, ニュース記事の文章のみの情報を利用する言語的アプローチとソーシャルメディアや WEB 上の情報といった外部の情報を利用するネットワークアプローチがある．

欺瞞的な文章には代名詞や接続詞の頻度や出現パターン，否定的な感情を表現する言葉が多く用いられるといった言語的な特徴が存在する [Feng 13]. 言語的アプローチはそうした特徴をニュースの文章から検出することを目的としている．

ソーシャルメディアで発信されたニュースに対して, ニュース記事の文章以外の特徴を用いてフェイクニュースの自動検出を行った例も存在する．[Tacchini 17] は Facebook の投稿がそれに「いいね！」を付けたユーザに基づいてデマであるかそうでないかに分類した．このように文書以外の情報を用いて分類する試みが存在するがフェイクニュースは必ずしも文書以外の情報が手に入るとは限らない．こういった手法は分類できる状況を限定してしまうという問題点がある．

連絡先: 山本和矢，yamamoto@complex.ist.hokudai.ac.jp

*1　https://www.cnn.co.jp/usa/35093206.html

3. 実験

3.1 使用手法

本実験では [Granik 17] が提案したナイーブベイズ分類器に基づくフェイクニュース検出器を実装して用いた．手法の詳細については [Granik 17] を参照されたい．

3.2 使用データ

先行研究で使われたデータは，入手不可能であったため代わりに形式の類似した別のデータを用いた．使用したデータは LIAR[*2] である．LIAR は [Wang 17] によってつくられたフェイクニュース検出用のベンチマークデータセットである．アメリカの政治にまつわる声明のファクトチェックを行っているウェブサイト POLITIFACT.COM[*3] の API で取得されたデータである．共和党及び民主党員の発言，ソーシャルメディアの投稿など約 1.2 万個の声明に対し，ID，6 段階の真偽ラベル (pants-fire, false, barely-true, half-true, mostly-true, true)，題目，発言者，発言者の職業，故郷の州，所属政党，発言者の声明に付けられたそれぞれのラベルの数，どこでの発言かといったタグが付与されている．実験では 6 段階のラベルのうち pants-fire, false, true のみを使用し，pants-fire と false をラベル Fake として扱った．またいくつかのデータの情報が欠損していたため除外した．

3.3 実験結果

ナイーブベイズ分類器を用いたフェイクニュース検出のデータの偏りによる再現率の変化を確認した．元のデータからラベルに偏りが生じるように抽出したデータとラベルに偏りがないように抽出したデータをデータ数の合計が同じになるように用意した．用意した 2 つのデータに対しナイーブベイズを用いたフェイクニュースの検出を行いそれぞれの再現率を比較した．偏りのあるデータの再現率は 0.04，偏りのないデータの再現率は 0.68 であった．これは [Granik 17] の仮説を支持する結果である．

Python の NLTK[*4] コーパスのストップワードリストに含まれる単語を辞書から除去した場合とストップワードを辞書から除去しなかった場合の精度，再現率，適合率，F 値を比較したところ，ストップワードの除去によってこれらが改善されず，むしろ僅かながら低下する傾向があった．これは，フェイクニュース検出においてはストップワードは除去すべきものではなく，むしろ検出に有効な情報を含んでいる可能性があることを示している．

辞書に含まれた 9328 種類の単語のうち，単語がニュース記事中に存在するときの，フェイクニュースである確率が大きい上位 30 単語を表 1 に示す．このうち socialists，muslim，bad について，これらの単語が出現する訓練データの文書の例を表 2 に示す．表から分かるようにこれらの単語は文書中において特定の人物や組織を批判する目的で使われている．このことからフェイクニュース検出の要因となった単語には特定の人物や組織を批判する目的で使われる単語が含まれていることが分かる．フェイクニュースらしさの大きい他の単語は，単語自体ではなぜそれがフェイクニュースと判定する要因とするのかが分かりにくいものが多い．しかし，たとえば responsible は is responsible for というフレーズで，他者を批判する文脈で用いられるなど，フレーズや文脈を考慮すれば，フェイクニュース判定の要因となった理由を解釈できるものもある．

∗2 https://www.cs.ucsb.edu/˜ william/ data/liar_dataset.zip
∗3 https://www.politifact.com
∗4 https://www.nltk.org/

face, governments, anything, responsible, began, reps, socialists, muslim, tell, scheme, cabinet, information, surplus, data, groups, census, benefit, reason, murphy, outside, loan, balanced,admits, parks, terry, provision, study, recently, bad, josh

表 1: フェイクニュース検出の要因となった単語

単語	単語が出現する訓練データの文書
socialists	Say Oregon Reps. Peter DeFazio and Earl Blumenauer are *socialists* who are openly serving in the U.S. Congress.
muslim	Says large majority of Republicans believe Obama is a *Muslim* and not U.S.-born.
bad	The Mexican government forces many *bad* people into our country.

表 2: 特定の単語が出現する訓練データの文書の例

4. まとめ

本稿では，先行研究で提案されたフェイクニュース検出の実装と評価を行った．これにより先行研究の仮説の検証と先行研究では着目されていなかったモデルの解釈性について分析を行いフェイクニュースに含まれる単語について知見を得ることができた．

本稿においては解釈性を活かすためナイーブベイズ分類器のみを用いたが，同様に解釈性のあるモデルに関して分類精度を比較する必要がある．検出の要因として単語だけでは解釈することが困難な場合が多いため，フレーズを抽出するなど，特徴のとり方を工夫するなどしてより人が解釈しやすくする必要がある．今回用いたフェイクニュース自動検出用のデータでは含まれる文書がアメリカの政治に関する限定的な題材であったが，様々なフェイクニュースを検出するためにはより多種多様なフェイクニュースに関するデータが必要であるため他の利用可能なデータの調査や SNS などを利用したデータの収集が必要である．将来的にはフェイクニュースの検出結果の根拠をユーザが解釈可能しやすい方法で可視化するシステムに関する研究を行いたいと考えている．

参考文献

[Feng 13] Feng, V. W. and Hirst, G.: Detecting deceptive opinions with profile compatibility, in *IJNLP*, pp. 338–346 (2013)

[Granik 17] Granik, M. and Mesyura, V.: Fake news detection using naive Bayes classifier, in *IEEE UKRCON*, pp. 900–903 (2017)

[Tacchini 17] Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., and Alfaro, de L.: Some like it hoax: Automated fake news detection in social networks, *arXiv preprint arXiv:1704.07506* (2017)

[Wang 17] Wang, W. Y.: ” liar, liar pants on fire”: A new benchmark dataset for fake news detection, *arXiv preprint arXiv:1705.00648* (2017)