解釈性のあるフェイクニュース検出器の実装に向けて

概要

フェイクニュース検出において、読み手が自身でニュースの信憑性を評価するために、解釈可能性は重要な要素である。我々は、解釈性のあるフェイクニュース自動検出器の実現に向けGranikら[1]の用いた手法を実装し実験を行った[2]。しかし、検出精度や、解釈可能性どちらも改善すべき点が多い。解釈性のあるフェイクニュースの実現に向けた現状の課題として、フェイクニュースをその内容のみから検出することの難しさ、多様なすべてのフェイクニュースを一つの手法で分類することの困難性、フェイクニュース検出における解釈可能性の定義が定まっていないことがあげられる。
それらを踏まえフェイクニュース検出の解釈としてどのような情報を提示すること妥当であるか、様々な方法を実装し実際に人手で評価を行い検証していく必要がある。

背景

フェイクニュース自動検出器には解釈可能性が必要
→検出結果が誤りであったり、検出器自体がフェイクである可能性を否定できないため

関連研究

“Fake news detection using naïve Bayes classifier.”(IEEE UKRCON 2017)
スパムメッセージとフェイクニュースの類似性を仮定し、スパムフィルターで用いられる**ナイーブベイズ分類器を使ってフェイクニュース検出**

それを踏まえて

先行研究

モデルの**解釈可能性に着目**
同じモデルを使ってフェイクニュース検出し、
検出の要因となった単語を分析

しかし

検出精度、結果の**解釈性**どちらも不十分

現状の課題

内容のみでの検出の困難性

- フェイクニュースの特徴が**本文から検出できるとは限らない**

フェイクニュースの多様性

- フェイクニュースには**多様な種類**がある

明確でない解釈性の定義

- 人がフェイクニュースだと**理解しやすい説明とは**

今後の展望

外部情報の利用

- ニュースの投稿者や題材の同じ他のニュースの情報を使う

ニュースの種類ごとに対応

- ニュースの種類ごとにその特徴をとらえた手法の使用

人手での検証

- 複数の説明手法を比較するため、人手で検証し評価

参考文献
[1] Granik, M. and Mesyura, V.: Fake news detec-tion using naive Bayes classifier, in *IEEE UKRCON*, pp.900–903 (2017)

[2]山本和矢, 小山聡, 栗原正仁. 解釈性のあるフェイクニュース検出器の実装と評価. 人工知能学会全国大会論文集 一般社団法人人工知能学会, pp. 3Rin237–3Rin237. 一般社団法人人工知能学会, 2019.