自然言語処理によるフェイクニュース判断の自動化

発表者: 総合情報学科 メディア情報学 コース 学籍番号 1510151 柳 裕太
指導教員: 田原 康之 准教授

1 はじめに

昨今の SNS の普及により，誰もが画像や動画と併せて情報を発信・収集できるようになった．逆に故意に情報を捏造して発信することによって，人々を誤った方向へ扇動するフェイクニュースも存在する．虚偽の情報ながら，扇動ではなく皮肉や風刺を込めたジョークニュースも存在する．ジョークニュースはフェイクニュースと同じく限りなく真実を模した形式をとるため，昨今では同じく SNS 上で拡散されやすく，同時に批判に晒されることもある．

本研究では，画像つきで発信された情報に対して，正しい情報か・フェイクニュースか・ジョークニュースかを判断するモデルを構築する．このモデルを使い，従来から画像・テキスト複合のデータセットに対して 3 カテゴリでも優秀な分類が行えることを示すことを目指す．それにより，SNS ユーザの情報収集を支援するエージェントの開発につなげることが可能となる．

2 関連研究

正しいニュース・フェイクニュース・ジョークニュースの 3 カテゴリ分類を機械学習で行う研究がある [2]．別の対象として，テキスト・画像を併せた情報を分類する機械学習モデルの検討も数多く行われている [3]．画像・テキスト双方を扱うモデルでは，実際に真実・フェイクとのカテゴリ分類において画像単独・テキスト単独の分類に比べて優秀な成績を収めていた [3]．しかしながら，あくまで“真実なのかそうでないのか”という 2 カテゴリで分類しているため，“他者を欺くための情報なのか，皮肉・風刺を込めた情報なのか”という観点での分析がなされていない．今回対象とする情報は，SNS 上で投稿された画像つきで発信されたニュースである．そのなかでも，正しいニュースを発信していたもの，フェイクニュースを発信していたもの，ジョークニュースを発信していたものが対象となる．

3 提案手法

3.1 モデル概観

今回提案したモデルは，以下の図 1 の通りである．

提案モデルの目的は，画像と文章で発信された情報に対して，正しいニュースか・フェイクニュースか・ジョークニュースかを分類するために，必要な特徴表現を学習することであった．提案モデルは複合特徴量抽出器とニュース分類器の大きく 2 部分に分けることができた．まず複合特徴量抽出器は，今回扱う情報が文章と画像を含むため，各メディアに対して特徴化する抽出器があった．その後それぞれの特徴を 1 つに連結し，複合特徴を形成した．複合特徴はニュース分類器に送られ，最終的には 3 カテゴリのどれに該当するかが判断された．

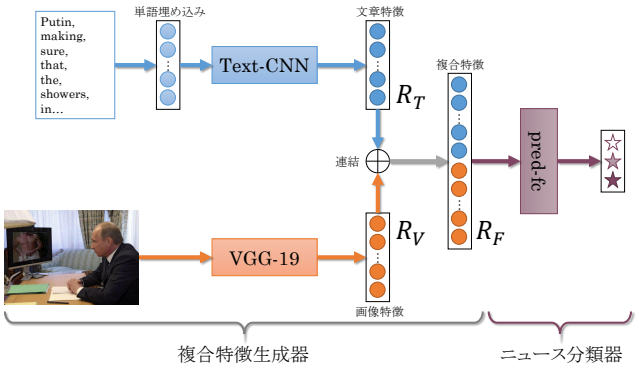

図 1: 提案モデル図．青: 文章特徴量抽出器，橙: 画像特徴抽出器，紫: ニュース分類器．

4 評価実験

4.1 実験条件

今実験では Twitter データセットを使用した [1]．

画像つき文章投稿を 3 カテゴリに分類する提案手法の有効性を調べるために文章のみで投稿を分類する手法 (以降，Text と表記)，もう 1 つは画像のみで投稿を分類する手法 (以降，Image と表記) を用意した．いずれも上記提案モデルから文章・画像特徴抽出器を除外したモデルを使用した．Text は入力データを提案モデルが使用したデータセットから画像を削除したものを使用し，Image は全投稿で使用された画像を対象とし，同じ画像に対して複数の文章投稿があった場合は 1 件として数えることにした．上記の条件を踏まえ，提案手法・Text・Image が扱う 3 カテゴリの投稿

件数は以下の表 1 の通りである．

表 1: 提案手法と比較対象手法が扱うカテゴリ毎の投稿数

手法	Real	Fake	Humor
Text	3021	4233	1509
Image	172	157	82
提案手法	3021	4233	1509

4.2 実験結果

3 モデルに対して 10 分割交差検定を行った結果が以下の表 2 の通りである．評価指標では，Precision(精度), Recall(再現率), F 値 (左 2 値の調和平均) のマクロ平均を使用することにした．

表 2: 各モデルの分類成果 (マクロ平均)

手法	Precision	Recall	F 値
Text	0.3649	0.3677	0.3016
Image	0.4942	0.5055	0.4667
提案手法	0.9268	0.9362	0.9286

この結果を見ると，提案手法が他 2 手法と比べて非常に高い分類成果を挙げたことが読み取れた．

5 評価

5.1 考察

今回の評価実験では，提案手法が 3 指標全てにおいて比較対象手法より優れた分類成績を収めた．これにより，SNS 上で画像つきの投稿を対象にした場合，正しいニュース・フェイクニュースの分類タスクのみならず，ジョークニュースも含めた分類においても従来のマルチメディアモデルのアプローチが有効であることが示唆されたのではないかと考えられる．

5.2 課題

今回分類するにあたり，大きな課題となったのが文章投稿の単語埋め込みへの変換であった．データセットが Twitter から収集されたものであったため，ユーザ生成コンテンツに対応することが難しかった．さらに，このモデルは英語のみを対象とした影響で，データセット内他国語投稿に対して対応ができなかった．

また，このモデルに限らずフェイクニュース検出というタスクにおいては，Wang らの研究 [3] によって問題点が指摘されていた．訓練データが扱うイベントや出来事の特殊性の影響を受けることにより，検証する時に訓練になかった別のイベントや出来事に対して正常な判断ができなくなる点であった．

6 おわりに

6.1 本論文のまとめ

本研究では，SNS 上で画像と文章を併せて発信された情報に対して，正しいニュース・フェイクニュース・ジョークニュースを判断するモデルを提案した．実際に 3 カテゴリ分類を行った結果，文章・画像単体から分類した場合に比べて，全ての評価指標において非常に優秀な分類成績を挙げた．これにより SNS 上における画像つき投稿に対して，ジョークニュースを含めた 3 カテゴリ分類も有効であることが示された．

6.2 今後の展望

このモデルの発展として，いくつかの方法が考えられる．

データセットが扱う出来事やイベントによる特殊性の対策として，Wang らの研究 [3] では敵対的生成ネットワーク (GAN) を模倣する形をとることが挙げられていた．真偽分類に加えて扱われたイベントも分類することによって，フェイクニュースの普遍的な特徴を抽出するようなアプローチが行われていた．

提案手法を日本語投稿に対応させることを考えた場合，残念ながら国内に今回使用したデータセットに近い規模をもつものがないため，SNS 上で日本語による画像つきの 3 カテゴリの投稿を収集する必要がある．

参考文献

[1] Christina Boididou et al. “Verifying Multimedia Use at MediaEval 2015.” In: *MediaEval*. 2015.

[2] Benjamin D. Horne and Sibel Adali. “This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News”. In: *CoRR* abs/1703.09398 (2017). arXiv: 1703.09398. URL: http://arxiv.org/abs/1703.09398.

[3] Yaqing Wang et al. “EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD ’18. London, United Kingdom: ACM, 2018, pp. 849–857. ISBN: 978-1-4503-5552-0. DOI: 10.1145/3219819.3219903. URL: http://doi.acm.org/10.1145/3219819.3219903.