平成 31 年度卒業論文

画像付きフェイクニュースとジョークニュースの検出・分類に向けた機械学習モデルの検討

電気通信大学 情報理工学部 総合情報学科
メディア情報学コース

学籍番号　：　1510151
氏名　：　柳 裕太

主任指導教員　：　田原 康之 准教授

指導教員　：　大須賀 昭彦 教授

指導教員　：　清 雄一 准教授

提出年月日　：　平成 31 年 2 月 8 日 (金)

概要

SNS の発展によりあらゆる情報入手が容易になった反面，人を欺くために故意に作成された虚偽の情報であるフェイクニュースが社会問題になっている．特に画像と併せて発信されたものは，テキストのみならず画像と併せた分析アプローチが有効である．虚偽の情報としては，もう 1 つジョークニュースというものもある．これは人を欺くためではなく，社会風刺や皮肉のために作られた情報という特徴がある．しばしばこの 2 カテゴリが混同され，ジョークニュースが批判に晒されることがあることも問題となっている．

既にテキスト・画像をニューラルネットワークの一種である CNN によって分析して真偽を判定する自動判別モデルが提案されている．実際に真実・フェイクとのカテゴリ分類において優秀な成績を収めているものの，ジョークとしての嘘情報と人を欺くための嘘情報が区別されていない．

本研究では，正しい情報・ジョークニュース・フェイクニュースの 3 カテゴリを分類することで，より画像つきフェイクニュースの検出精度を向上させることを目指した．

実際に SNS から収集した画像つきのデータセットを対象にカテゴリ分類を行った結果，3 カテゴリでもマクロ F 値が約 0.93 と良好な結果を示した．

目次

第 1 章

序論

1.1 背景

昨今の SNS の普及により，誰もが情報を発信・収集できるようになった．特に最近ではテキストのみならず，画像や動画と併せて情報の発信が可能である．一般論として，テキスト単体と比べて画像や動画と併せて発信されたマルチメディア情報の方が多くの注目を得やすい．逆にこれを利用して，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．フェイクニュースが広まると、大規模なマイナスの影響が出る可能性があり、場合によっては重要な公共の出来事に影響を及ぼしたり、操作したりすることさえある．例えば 2016 年の米国大統領選では，2 名の候補者を支持させるためのフェイクニュースが多く拡散され，とりわけ Facebook 上では 3700 万回以上共有された [1].

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．有名な発信メディアとしては，英語では the Onion，日本語では虚構新聞が該当する．あくまで扇動ではなく笑いを提供するためのものであり，多くの場合それは批判の的にはなりにくい．しかしながら，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，同じく SNS 上で拡散されやすい傾向にある．

当研究では，扇動のために故意に情報を捏造して発信された情報をフェイクニュース，事実を発信した情報を正しいニュース，そして風刺や皮肉を込めて発信された情報をジョークニュースとして定義する．

1.2 先行研究

フェイクニュースに限らず，風評や web ページの信憑性を評価するモデルの構築の研究は数多く行われている．例えば，福島らの研究 [17] では，web ページの体裁から信頼性を評価するモデルが提案された．また，機械学習による分類が非常に盛んに行われている．なかでも Granik らの研究 [5] や Gilda の研究 [4]，そして松尾の研究 [16] により，単語埋め込みと

ナイーブベイズ分類器や SVM，決定木といった教師あり学習を組み合わせることによって，フェイクニュースや流言を分類するタスクで優秀な分類成果を挙げることが報告された．ほかにも Wu らの研究 [15] によると，SNS 上で拡散された情報に対して，“誰が・どのような経緯で拡散したか” という情報から信憑性を判断するモデルも提案された．Rubin らの研究 [11] によれば，正しいニュース・ジョークニュースの分類にもこのアプローチが有効であることが示されていた．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 カテゴリ分類においても研究が行われている．特に Horne と Sibel の研究 [6] によると，フェイクニュースは正しいニュースよりジョークニュースに近い性質をもち，真実に近い形式をとるほど高い説得力をもつことが示されていた．

上記の機械学習を使った研究では，いずれもテキストのみの情報を対象としていた．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．大まかな形としては，まずテキスト・画像を何らかの方法でベクトル化する．その後 2 種のベクトルを結合し，真偽判定を行うモデルに渡す形をとっている．例えば Jin らの研究 [7] では，テキストでは LSTM，画像では VGG-19 を使用してベクトル化しており，更に Attention とソーシャルコンテキスト (ハッシュタグ，URL 等) によって更に高精度な分類を行うモデルが提案されていた．また Wang らの研究 [14] では，EANN というモデルが提案されている．これは画像のベクトル化においては同じく VGG-19 を使用しているが，テキストではテキスト CNN を使用していた．

1.3　研究課題

上記の EANN モデルのような画像・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 [14]. 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．

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．

上記の提案する情報分類システムを検証するために，事前に用意されたデータセットを用いて 10 分割交差検定によって分析を行った．また上記システムの分類性能を評価するために，画像・テキスト単独で分類を行った結果と比較することで，提案システムが目標に適していることを示すことを目指した．その結果テキスト単独でのマクロ F 値が約 0.22，画像単独でのマクロ F 値が約 0.40 であったのに比べ，提案モデルのマクロ F 値は約 0.93 という数値を出

し，提案モデルの有効性が示された．

第 2 章

提案手法

2.1　モデル概観

この章では，提案モデルがもつ複合特徴量生成器とニュース分類器について紹介する．その後にこの 2 要素を統合して転移学習が可能な表現を学習する方法について説明する．今回提案したモデルは，以下の図 2.1 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生成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．まず複合特徴量生成器は，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生成器があった．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．

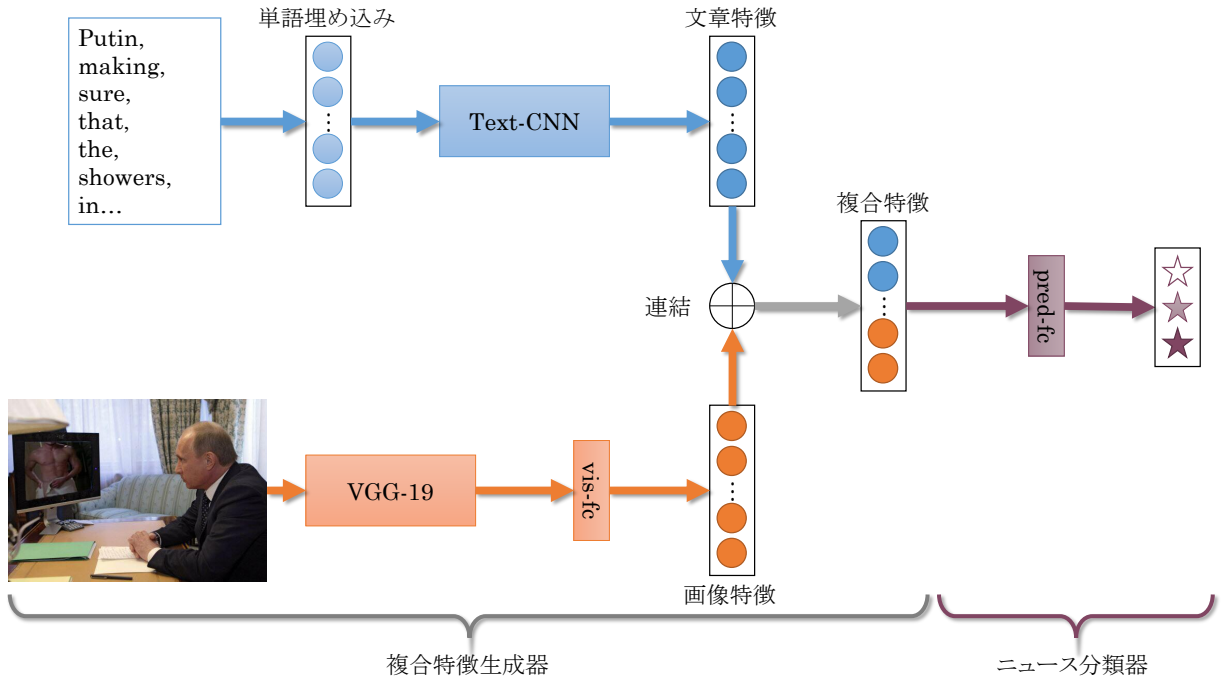

図 2.1　提案モデル図．青色が文章特徴量生成器，橙色が画像特徴生成器，紫色がニュース分類器であった．

2.2　複合特徴生成器

2.2.1　文章特徴

文章特徴は，入力に英語の投稿をスペース毎に分割した英単語の連続リストをもった．まずは単語を単語埋め込みでベクトル化した．その後単語の羅列から分類に有効な情報を得るために，文章特徴を生成する核として CNN (convolutional neural networks: 畳み込みニューラルネットワーク) を採用した．CNN はコンピュータビジョンやテキスト分類などの多くの分野で効果的であることが示されていた [3, 8]．図 2.1 の通り，提案手法では CNN の発展形であるテキスト CNN(Text-CNN)[9] を採用した．テキスト CNN の構造は図 2.2 の通りである．複数のウィンドウで畳み込むことで，様々な角度から特徴を抽出することを実現した．

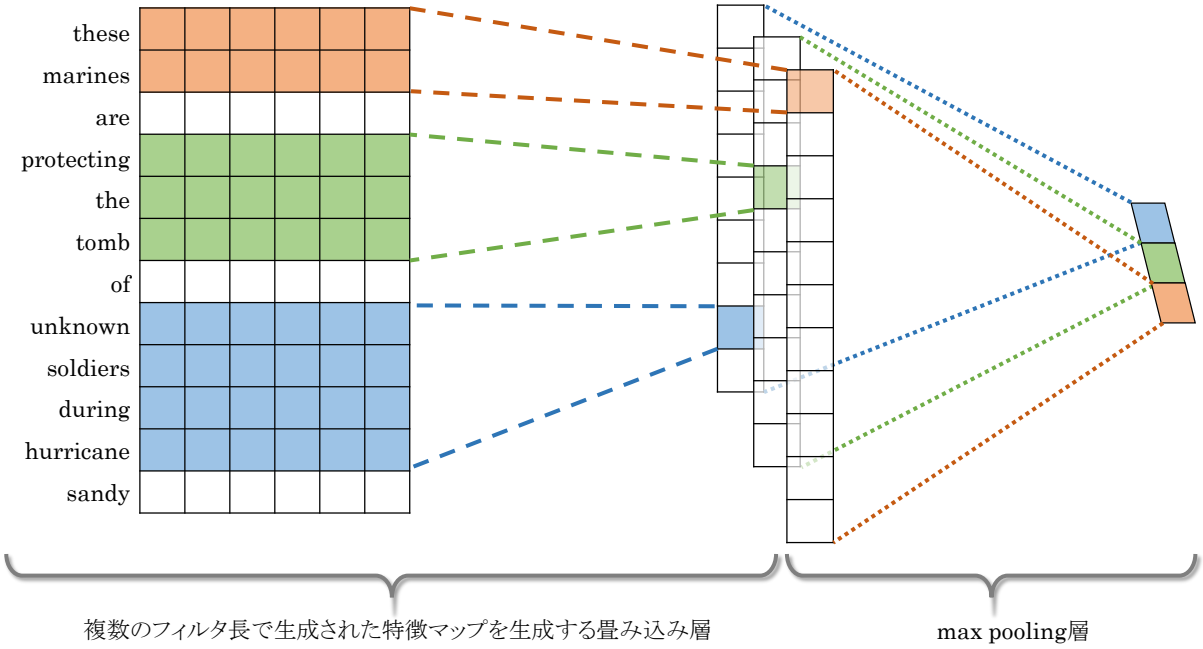

図 2.2　テキスト CNN の図．Wang らの研究 [14] を参考に作成．

具体的な手法では，EANN が採用したテキスト CNN と全く同じ流れを採用した [14].

2.2.2　画像特徴

画像から効率的に特徴を生成するために，当研究では事前学習済みの VGG-19[12] を起用した．VGG-19 は畳み込み 13 層と全結合層 3 層から形成され，最終的には 1000 次元の特徴ベクトルが出力された．当研究では最終の全結合層のみ改変し，文章特徴のベクトル次元数と同じ数の次元をもつベクトルを出力されるようにした．また改変した最終全結合層以外は，過学習を防ぐために事前学習の状態を維持することにした．

こうして文章特徴・画像特徴が生成され，最終的には 2 つの特徴ベクトルを 1 つに連結したものが複合特徴となった．

2.3　ニュース分類器

複合特徴はニュース分類器にて正しいニュース・フェイクニュース・ジョークニュースとして分類された．具体的には隠れ層を含む全結合層と softmax から形成され，最終的な分類が行われた．

第 3 章

評価実験

3.1　データセット

今回の実験での訓練データセットでは，Boididou らの研究 [2] によって提案された Twitter 投稿データセットを使用した．こちらも Twitter 上でフェイクニュースを検出するために作られたデータセットであるが，付加されたラベルとして Real, Fake, そして Humor があり，ジョークニュースを含めた 3 カテゴリ分類に適したものとなっているため，当研究で採用した．データセットでは訓練用と検証用として 2 部に分かれていたが，当研究では訓練用とされた部分を対象に 10 分割交差検定することにした．データセット内ではツイート文章と画像のみならず，タイムスタンプや投稿者といったソーシャルコンテキスト情報も含まれている．当研究ではソーシャルコンテキストは対象に含まず，文章と画像のみで 3 カテゴリ分類することを目指した．

3.2　比較対象手法

今回，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 2 種類の比較対象手法を用意した．1 つは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) であった．いずれも上記提案モデルから文章・画像特徴生成器を除外したモデルを使用した．また 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た．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．

3.3　実験条件

3.3.1　使用データ統計

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 3.1 の通りである．Text が使用するデータは提案手法が扱うデータから画像を削除したものであるため，提案手法と全く同じ件数になった．Image は，同じ画像に対して複数の投稿があったため，他 2 手法と比べて非常に少なくなっている．

表 3.1　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

3.3.2　モデル条件

Text

まず単語埋め込みに変換する際，Google News データセットから事前学習済みの word2vec モデル [10] を使用した．このモデルでは，各単語を 300 次元のベクトルに変換するものであった．ここで wod2vec モデルに該当しない単語が出現した場合，<unknown>として seed 値固定ランダムベクトルを生成することにした．また，投稿の全単語中 50% 以上が<unknown>の場合，実態に則さない学習を避けるために学習対象から外すことにした．その後テキスト CNN に送られ，1 つの文章に対して 1 つの 300 次元のベクトルが生成され，ニュース分類器に渡す形となった．なお，フィルタサイズは 2-5 までとし，隠れ層は 1 つ用意し，ユニット数は 300 とした．隠れ層では 60% のユニットが無視される Dropout を導入した．ニュース分類器内でも隠れ層は 1 つ用意し，ユニット数は 300，上記と同じ条件の Dropout も導入した．

Image

モデルの複雑化を避けるため，1 つの投稿に複数枚画像が付加されていた場合は最初の 1 枚のみをモデルに入力させることにした．画像は事前学習済み VGG-19 モデルに入力し，1 つの画像に対して 1 つの 300 次元のベクトルが生成され，ニュース分類器に渡す形となった．本来の VGG-19 は最終層にて 1000 次元のベクトルが出力されるが，最終層のみ改変して 300 次元のベクトルが出力されるようにした．また前記の通り過学習を避けるために最終層を除き事前学習済みの状態を維持させることにした．ニュース分類器の条件は上記 Text と同一で

あった．

提案手法

提案手法では，Text と Image を統合した形をとったため，画像・文章の部分は上記と同様の条件をとった．画像・文章の特徴を結合するため，ニュース分類器に渡されるのは 600 次元のベクトルであった．それにあわせ，隠れ層のユニット数も 600 とした．

3.3.3　評価指標

評価指標では，Precision(精度), Recall(再現率), F 値 (左 2 値の調和平均) を使用することにした．算出する方法上各カテゴリ毎に上記指標があるが，今回使用するデータセットでは極端にカテゴリが偏っていないので，各カテゴリの指標を先に算出してから 3 カテゴリの平均をとるマクロ平均を評価に使うことにした．

3.4　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 3.2 の通りである．

表 3.2　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．また，比較対象手法内で比べると画像単体の方が分類成果が良好である点もみられた．

第 4 章

評価

4.1 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．

また比較対象手法に限って結果を観察すると，文章単体より画像単体の分類の方が優秀な分類成績であった．これは自然言語より画像の方が分類タスクにおいて研究が進んでいることや，SNS 上の投稿であった故に単語埋め込みに変換する際に<unknown>に変換されやすい傾向にあったことや，文章の場合英語以外の投稿に対応できないものの，画像においては英語圏以外の投稿であっても十分言語の違いに影響されにくかったことなど，いくつかの原因が推察される．

4.2 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．例えば今回使用したデータセットが Twitter から収集されたものであったため，事前学習済み word2vec モデルが対応できない短縮語や造語 (ハッシュタグなど) といったユーザ生成コンテンツに対応することが難しかった．

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 [14] によってある問題点が指摘されていた．訓練に使ったデータセット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が使われた場合に正常な判断ができなくなる点であった．

さらに，このモデルは英語のみを対象としたものであった点も挙げられた．データセット内一部では他国の言語が含まれていたため，単語埋め込みに変換する際に大幅に<unknown>に変

換される傾向もあった.

第 5 章

おわりに

5.1　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．

5.2　今後の展望

このモデルの発展形として，いくつかの方法が考えられる．

例えば文章特徴生成器に対して，テキスト CNN ではなく Vosoughi らの研究 [13] によって SNS 投稿を分析するために提案された，文字単位でベクトル変換する方法を採用することなどが考えられる．

また，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 [14] 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．このイベントや出来事による特殊性を排するために，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，特徴化する際に特殊性を排し，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．実際にこれによって分類精度が改善した点が上記研究によって報告されていたため，当研究でも有効に働く可能性がある．

提案手法を日本語投稿に対応させることを考えた場合，まず SNS 上で日本語による画像つきの正しいニュース・フェイクニュース・ジョークニュースの投稿を収集する必要があると考えられる．その上に形態素解析によって分かち書きする部分を加え，さらに日本語用の事前学習済み word2vec を用意する必要だと考えられる．もしも既に日本語投稿による 3 カテゴリ分類済みのデータセットがあれば投稿を収集する必要はないが，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のが現状である.

謝辞

本研究を行うにあたり，ご多忙の中，終始適切かつ丁寧なご指導をして下さった大須賀昭彦教授，田原康之准教授，清雄一准教授に深く感謝致します．貴重な勉学の機会を与えてくださったことに深く御礼申し上げます．

参考文献

[1] Hunt Allcott and Matthew Gentzkow. “Social media and fake news in the 2016 election”. In: *Journal of Economic Perspectives* 31.2 (2017), pp. 211–36.

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

[3] Ronan Collobert et al. “Natural language processing (almost) from scratch”. In: *Journal of Machine Learning Research* 12.Aug (2011), pp. 2493–2537.

[4] S. Gilda. “Evaluating machine learning algorithms for fake news detection”. In: *2017 IEEE 15th Student Conference on Research and Development (SCOReD)*. Dec. 2017, pp. 110–115. DOI: 10.1109/SCORED.2017.8305411.

[5] M. Granik and V. Mesyura. “Fake news detection using naive Bayes classifier”. In: *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*. May 2017, pp. 900–903. DOI: 10.1109/UKRCON.2017.8100379.

[6] 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: *arXiv preprint arXiv:1703.09398* (2017).

[7] Zhiwei Jin et al. “Multimodal fusion with recurrent neural networks for rumor detection on microblogs”. In: *Proceedings of the 2017 ACM on Multimedia Conference*. ACM. 2017, pp. 795–816.

[8] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. “A convolutional neural network for modelling sentences”. In: *arXiv preprint arXiv:1404.2188* (2014).

[9] Yoon Kim. “Convolutional neural networks for sentence classification”. In: *arXiv preprint arXiv:1408.5882* (2014).

[10] Tomas Mikolov and Ilya Sutskever. *Google Code Archive - Long-term storage for Google Code Project Hosting.* July 2013. URL: https://code.google.com/archive/p/word2vec/ (visited on 01/24/2019).

[11] Victoria Rubin et al. “Fake news or truth? using satirical cues to detect potentially misleading news”. In: *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*. 2016, pp. 7–17.

[12] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[13] Soroush Vosoughi, Prashanth Vijayaraghavan, and Deb Roy. “Tweet2Vec: Learning Tweet Embeddings Using Character-level CNN-LSTM Encoder-Decoder”. In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’16. Pisa, Italy: ACM, 2016, pp. 1041–1044. ISBN: 978-1-4503-4069-4. DOI: 10.1145/2911451.2914762. URL: http://doi.acm.org/10.1145/2911451.2914762.

[14] 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*. ACM. 2018, pp. 849–857.

[15] Liang Wu and Huan Liu. “Tracing fake-news footprints: Characterizing social media messages by how they propagate”. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM. 2018, pp. 637–645.

[16] 松尾省吾. “機械学習を用いた流言の検出に関する研究”. MA thesis. 電気通信大学院, 2018.

[17] 福島隆寛 and 内海彰. “Web ページの信頼性の自動推定”. In: **知能と情報** 19.3 (2007), pp. 239–249.