数理解析研究所講究録 第2124巻 2019年 17-27

An Empirical Study on Fake Review Detection by Latent Dirichlet Allocation

株式会社ファンコミュニケーションズ情報科学技術研究所 崎濱栄治

F@N Communications, Inc. Information Science and Technology Institute, Eiji Sakihama

青山学院大学経営学部 鹿島浩之

Aoyama Gakuin University Department of Business Administration, Hiroyuki Kashima

概要

インターネットを利用した購買行動は身近なものとなった。多くの EC サイトではユーザーの利便性を高めるために、レビューの投稿や総合評価を記録し比較検討の材料を提供している。一方、レビューや総合評価を意図的に操作するなど不正も問題となっている。本研究では、レビューのタイトルと本文、総合評価からフェイクレビューを特定し、トピックモデルの一つである Latent Dirichlet Allocation(LDA) による特徴量を推定した後、ランダムフォレスト によるフェイクレビュー判定モデルの検証を行った。単純な文字数情報によるモデルと比較して、LDA によって推定された特徴量を利用することで大幅に予測精度が向上した。また、フェイクレビューと関連の深いトピックは特徴量の感度分析から、予測に対して有益であることが確認された。

1 序論

1.1 背景と問題意識

商品やサービスを比較検討する際にレビューや総合評価は有益な情報であるが、利害関係者等によるフェイク (偽) レビューによる過大 (過少) 評価が問題となっている [10]Mukherjee,et al.(2012) 等)。例えば、自動プログラムを利用して機械的にポジティブなレビューの投稿を行うことで、意図的に利害関係のある商品やサービスが検索されやすいようにする等の例がある。レビューに対してユーザーが投票する仕組みが実装され、ある程度有用性が担保されているケースもあるが、一般ユーザーによる投票と悪意のあるユーザーやボットと呼ばれるプログラムによる投票を見分けることは困難である。信頼に値しないレビューを客観的に評価し、排除することができれば、ユーザーの利便性は大きく向上するであろう。レビュー本文のサンプルを図 1 に示す。

```
[1] "暇つぶしにいい!ちょっと考える問題もあるけどヒントを見ればたいがい解けます"
[2] "気分がいい非常に良い英語学習ソフトウェアは、懸命に作業を続けることを願っています!"
[3] "国際専門アプリストア最適化(ASO)1、AppStoreキーワード検索ランキングをトップ3に上げられる
、もしトップ3にならなかったら無料にする2、AppStore五スター好評。Appのキーワード関連数が増やす
```

図 1 レビュー本文のサンプル

1 番目のレビューは、カジュアルなクイズかパズルゲームに関するレビューであることが読み取れる。2 番目は、英語学習に関するアプリであることが示唆されるが、日本語としてやや不自然な文章である。3 番目についても不自然な文章である。なお、3 番目については特定の SNS の ID がレビュー本文に含まれていたため当該箇所は削除している。このようなレビューが投稿される背景に、レビュワーが何らかの報酬を得られる仕組みの存在が推測される。一般のユーザーがこのようなレビューから有益な情報を得ることは、ほぼ期待できずスパム／フェイクレビューの一種であるとみなすべきであろう。

1.2　先行研究

[13]Igi,et al.(2014) は、レビューの信頼性を表す指標として、類似性、協調性、集中性及び、情報性という 4 つの信頼性指標を定義し、各指標ごとのスコアを求め、そのスコアを可視化して提示した。[14] 中里,et al.(2014) は、レビュー投稿の時系列情報と総合評価の点数を元に、レビューワーとレビューの信頼性を評価した。[17] 三舩,et al.(2016) は、[9]Ott,et al.(2011),[13]Igi,et al.(2014) をベースにフェイクレビューをルールベースで特定し、名詞/形容詞についてランダムフォレストでフェイクレビュー分類時に重要な特徴語を抽出した。これらの研究では、レビュー文章が持つ潜在的な意味については未活用である。

[19] 高島,et al.(2017) は、レビュー投稿が参考になった場合に投票される「Like」数を有用度と定義し、レビュー文書の構造/統語/意味の 3 カテゴリの特徴量を用いてサポートベクター回帰による有用性を判定した。有用性の高いレビューについては示唆が得られる一方、フェイクレビュー判定は未着手である。[23] 岡山,et al.(2018) は、フェイクニュースで学習した SVM による分類器でフェイクレビューの分類を行った。フェイクラベルは所与であること、TF-IDF+PCA による次元削減結果を特徴量としており、潜在意味的な特徴量はなく、英文が対象であった。

1.3　研究の目的

本研究では、日本語のレビューを対象とし、データの入手が容易なレビュー文章とタイトル、総合評価のみでフェイクレビューの特定を行う。その際に、Latent Dirichlet Allocation(LDA) を活用することでレビューが内包する潜在トピックを抽出し、フェイクレビューと関連の深い潜在トピックの推定を試みる。潜在的な意味を考慮することで表記揺れに対応した頑健性が期待される。

アプリレビューの特徴として、アプリのジャンル、不具合や改善要求、便利や有益といった感想、広告の有無、利用料金、使用時の通信量など複数の話題（トピック）が存在すること、略称や表記揺れ (例：引っ越し/引越し/引越、いぬ/イヌ/犬など) が多いことが挙げられ、これらに対応した手法を採用する必要がある。

2　分析の枠組み

本研究では、フェイクレビューの特定、LDA に適合したデータの作成、LDA の適用、ランダムフォレストによる予測モデルの構築と結果の検討を行う。

以下、本研究の流れを示す。

1. フェイクレビューの特定
2. LDA におけるトピック数の検討
3. LDA によるトピック確率の推定

4. ランダムフォレストによるフェイクレビュー予測モデルの検討

実装は R で行い、LDA については topicmodels パッケージ、ランダムフォレストについては、randomForest パッケージを利用した。

2.1 データの作成方法

アプリのレビュー文書を対象として各レビュー文書を形態素解析し単語リストに変換する。形態素解析には RMeCab を利用した。レビュー文書には新語やネット特有の言い回しが多用されるため新語/固有表現に強いとされる辞書 (NEologd) を利用した [21]。単語リストに含まれる単語は形容詞、名詞のみとし、ストップワード (一般的すぎるため分析から除外すべき単語) については SlothLib のリスト[*1]を取得し削除した。得られた単語リストを文書単語行列に変換し、LDA モデルを適用することでトピックを抽出する。

2.2 Latent Dirichlet Allocation(LDA)

[22]Katsumata,et al(2017) に即し、LDA による文書生成過程を説明する。M 個のレビュー文章は、V 個のボキャブラリー (単語の種類) から作られ、文書全体の背後に K 個のトピック

$$\phi_k = (\phi_{k,1}, ..., \phi_{k,V}) \quad (k = 1, ..., K)$$

が、個々の文書 d の背後には各トピックの出現確率分布

$$\theta_d = (\theta_{d,1}, ..., \theta_{d,K}) \quad (d = 1, ..., M)$$

がそれぞれ存在しているものとする。ここで、$\phi_{k,v}$ はボキャブラリー中 v 番目の単語の出現確率、$\theta_{d,k}$はk番目のトピックの出現確率であり、以下を満たす。

$$\sum_{v=1}^{V} \phi_{k,v} = \sum_{k=1}^{K} \theta_{d,k} = 1$$

文書 d を構成する単語 $w_{d,i} \in \{1, ..., V\}\,(i = 1, ..., n_d)$ は、これに対応したトピック $k_{d,i} \in \{1, ..., K\}$ が背後に存在し、

$$w_{d,i} \sim Cat_V(\phi_{k_{d,i}}) \quad (i = 1, ..., n_d)$$

によって出現するものとする。ここで、n_d は文書 d の単語数、Cat は Categorical 分布 (1 回試行の多項分布) である。そして、$k_{d,i}$ は、

$$k_{d,i} \sim Cat_K(\theta_d) \quad (i = 1, ..., n_d)$$

により与えられているものとする。更に、Categorical 分布のパラメータは、それぞれ以下の Dirichlet 分布から生成されているものと仮定する。

$$\begin{aligned} \phi_k &\sim Dir_V(\alpha) \quad (k = 1, ..., K), \\ \theta_d &\sim Dir_K(\beta) \quad (d = 1, ..., M) \end{aligned}$$

[*1] http://svn.sourceforge.jp/svnroot/slothlib/CSharp/Version1/SlothLib/NLP/Filter/StopWord/word/Japanese.txt

ここで、$\alpha = (\alpha_1, ..., \alpha_V)$、$\beta = (\beta_1, ..., \beta_K)$ は、Dirichelet 分布のパラメータである。レビュー文書は、上記のモデルから生成されるものと仮定する。このとき観測データから Collasped Gibbs Sampling を経て、ϕ_k、θ_d の推定が可能となる。(詳細は [22]Katsumata,et al.(2017) 参照)。

2.3 ランダムフォレスト

Random Forests は、複数の木 (tree) を用いて (forest) を構成して識別などを行う機械学習アルゴリズムである。個々の決定木は高い識別性をもつわけではないが、それらを複数用いることによって高い予測性能を得るという特徴がある。(詳しくは [11] 波部斉.(2012) を参照)。ランダムフォレストは過学習を避けつつ、外れ値にも強い好ましい性質を持ち、特徴量の重要度や感度分析の実行も可能であることから、研究、実務共に広く活用されている。

3 実証分析

3.1 データ概要

App Store(2008 年 7 月 10 日～2018 年 11 月 7 日) から日本国内対象のアプリについて、レビューのタイトル、本文、Rating(総合評価) の約 360 万件を取得した。

3.2 フェイクレビューの特定方法

本研究においては、明確な教師データとしてフェイクラベルを付与することが不可能なため [5]Jindal,et al.(2007)、 [10]Mukherjee,et al.(2012) を参考とし、下記条件を満たした場合にフェイクレビューとした。

- レビューの文字数が 20 文字以上
- rating が最小 (1) or 最大 (5)
- 複数レコード存在する

3.3 データ分割

フェイクレビューラベルを特定した後、全データから訓練データとテストデータのサンプリングを行った。

表 1　訓練データとテストデータの内訳

ラベル	訓練データ	テストデータ
non-フェイク	42,312	18,083
フェイク	3,199	1,455
合計	45,511	19,538
フェイク比率	7.03%	7.45%

4　分析結果と考察

4.1　トピック数の決定

文書生成モデルを評価する指標としては、広くパープレキシティが利用されている [7]。処理対象となる文書 D_{test} の総数を M とした場合、パープレキシティは以下の式 (1) で計算され、値が小さいほど予測精度が高いと考えられる。

$$Perplexity\left(D_{test}\right)=\exp\left\{-\frac{\sum_{d=1}^{M}\log p\left(d_d\right)}{\sum_{d=1}^{M}N_d}\right\} \tag{1}$$

(N_d は文書 d の長さ、$p(d_d)$ はモデルによって文書 d が生成される確率。)

トピック数の決定に関しては複数の指標が提案されている。本研究においては複数の観点からトピック数を評価するため Perplexity の他に潜在トピック間のコサイン類似度に注目した Griffiths2004、対数尤度に注目した CaoJuan2009 の指標についても検討対象とした。Perplexity はトピック数=100 が最も小さい値となったが、50 との差は非常に小さかった。Griffiths2004 と CaoJuan2009 でもトピック数=50 と 100 で性能に大きな差はなかった。経験的にトピック数=100 は多すぎること、3 指標で見てトピック数=50 と 100 で大きな差がなかったことからトピック数=50 を採用した。

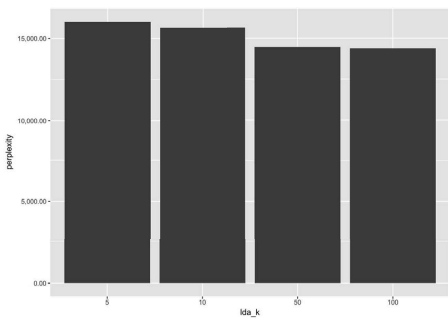

図 2　Perplexity

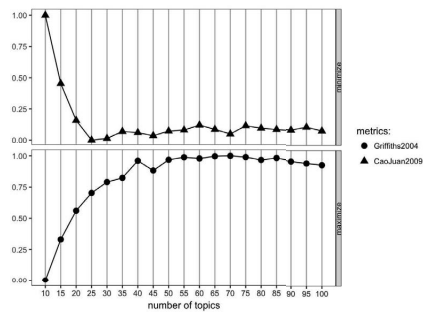

図 3　Griffiths2004,CaoJuan2009

4.2　得られたトピックの考察

本分析によって得られたトピックの一例を示す。

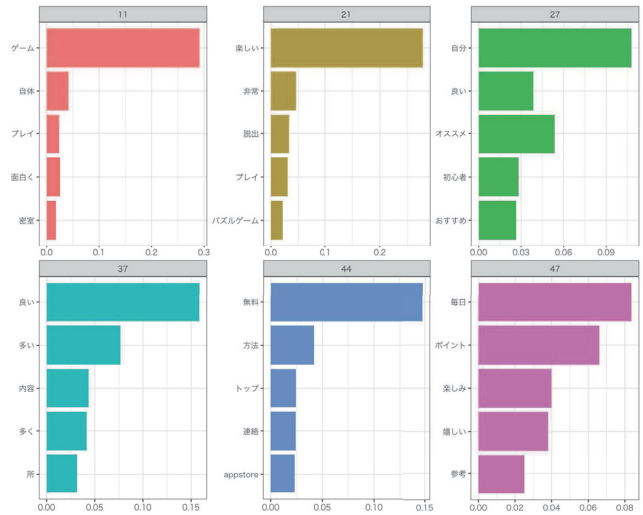

図 4　トピック抽出結果の一例

トピック 44 は図 1 のサンプルで示したレビュー本文例の 3 番目に含まれる「無料」「方法」「トップ」といった単語の出現確率が高く、同トピックの確率が高いレビューはフェイクの可能性があるといえる。トピック 21 は「楽しい」の出現確率が最も高く、ポジティブなレビューであることが示唆されているが、「脱出」「パズルゲーム」といったゲームのジャンルに関する単語の出現確率が高い。よって特定のゲームについて、ポジティブなレビューが集中したことが示唆されておりフェイクレビューの可能性も疑われる。

4.3　ランダムフォレストによるフェイク予測

ランダムフォレスト によるフェイクラベルの分類器の検討を行う。

4.3.1　検討モデル

model-1:文字数のみ

$$label_{fake} : length_{title} + length_{review} \tag{2}$$

model-2:文字数 + トピック確率

$$label_{fake} : length_{title} + length_{review} + topic_1 \sim \ldots \sim + topic_{50} \tag{3}$$

本研究ではレビュータイトル、レビュー本文の文字数のみを特徴量とした model1 の (2) 式と、**model-1** に加えて、LDA の結果得られるレビュー毎のトピック確率 ($topic1...topic50$) を考慮した **model-2** の (3) 式を検討する。

4.3.2　予測精度

model-1、**model-2** の予測精度を再現率 (実際にフェイクであるもののうち、フェイクであると予測され

たものの割合) で比較する。表 1 のとおり本研究のような不均衡データの場合、全て non-フェイクと予測したとしても訓練データでは 93.97% 、テストデータでは 93.55% の正解率となってしまうためである。

表 2　**model-1** の予測精度 (再現率 = 8.52%)

		真値	
		non-フェイク	フェイク
予測値	non-フェイク	18,058	1,331
	フェイク	25	124

表 3　**model-2** の予測精度 (再現率 = 27.01%)

		真値	
		non-フェイク	フェイク
予測値	non-フェイク	18,061	1,062
	フェイク	23	393

model-1 の再現率は、8.52% と比較して、**model-2** の再現率は 27.01% と大幅に改善していることから LDA によるレビュー文書のトピック確率はフェイク予測に対して有効であることが示唆された。

4.3.3　特徴量の重要度

ランダムフォレストの結果からジニ係数の平均的な減少量に基づいた特徴量の重要度が得られる。

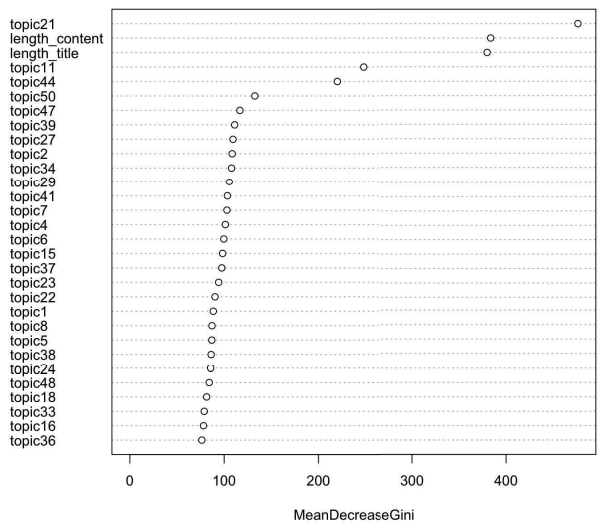

図 5　特徴量の重要度

最も重要度の高い特徴量は図 4 で確認したトピック 21 であった。次に、レビュー本文の文字数、レビュータイトルの文字数、トピック 11、トピック 44 と続いている。

4.3.4　特徴量の感度分析

ランダムフォレスト の予測において特定した重要な特徴量について感度分析を Partial Dependence Plot によって行う。簡潔にまとめると、その他の条件を一定とし注目する特徴量のみを変化させた場合に、フェイクと判定される確率がどのように変化するか確認することができる。

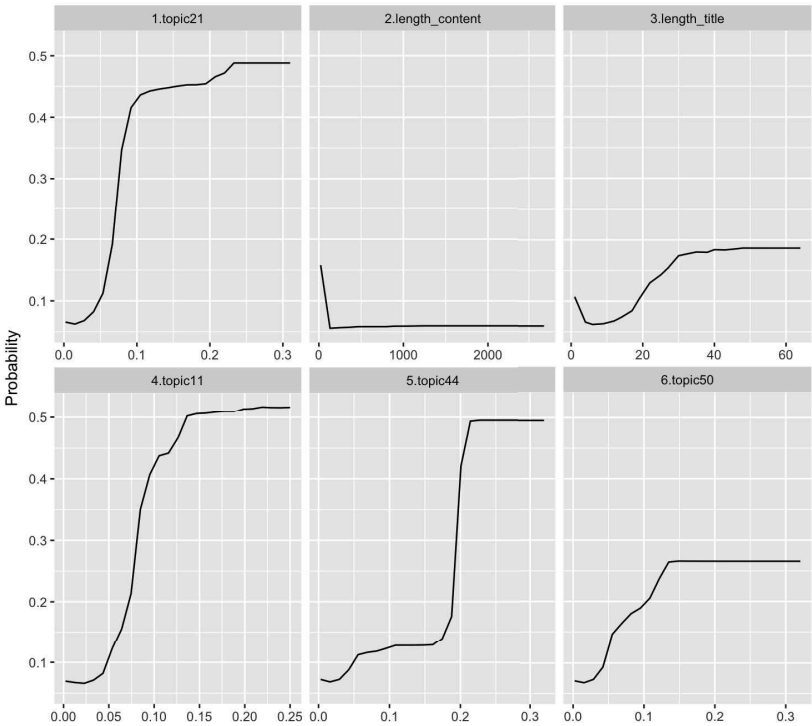

図 6　感度分析 :Partial Dependence Plot

重要度の最も高かったトピック 21 について確認すると、値が 0.05 を越えるとフェイクと判定される確率が急上昇し、その後傾きは緩やかになる。その他のトピックについてもある範囲では急激に上昇し、緩やかになる傾向が確認できた。レビュー本文の文字数は約 100 文字以上になるとフェイクの判定確率は 0% に近くなっているが、フェイクレビューの特定は 3.2 のとおり 20 文字以上の文字列一致が条件となっていることから、文字数が多くなるとフェイクレビューであると判定されにくくなることが想定される。レビュータイトルについては文字数が 5 文字より少なければフェイクレビューとされる確率が 5 – 10% 程度あるが、増加するに従って 20% を上限に上昇している。

4.3.5　レビューの例 1(予測値=フェイク、真値=フェイク)

予測精度で確認した表 3 の中で、予測値がフェイクかつ真値もフェイクであった 393 件から例をあげる。

表 4 におけるレビュー本文の 1,2 番目は日本語として不自然な文章であり、どちらも表 5 からトピック 44 の値が 0.2 を超えており、図 6 の感度分析からはトピック 44 の値だけでフェイクレビューの予測確率が 50% となることが示唆されている。3,4,5 番目のレビューについても日本語としてはやや不自然であり、特定のアプリに集中して高評価のレビューを付与していたことがうかがわれる。これらのレビューは、図 4 で確認したトピック 21 の値が高かった。

表 4　レビュー本文

No	レビュー
1	国際専門アプリストア最適化 (ASO)、AppStore キーワード検索ランキングをトップ 3 に上げられる、 もしトップ 3 にならなかったら無料にする、AppStore 五スター好評。App のキーワード関連数が増やす連絡方法……
2	AppStore キーワード検索ランキングをトップ 3 に上げられる、もしトップ 3 にならなかったら無料にする AppStore 五スター好評。App のキーワード関連数が増やす連絡方法……
3	ああ止めることができなかっただけで脱出げーむ！
4	グッドデザイン、良いパズル脱出パズルゲームを開発しています
5	ファーストプレイ脱出パズルゲーム、非常にエキサイティング

表 5　特徴量

No	*topic*21	*topic*44	予測確率
1	0.014	**0.319**	100%
2	0.014	**0.289**	100%
3	**0.091**	0.018	92%
4	**0.190**	0.015	98%
5	**0.174**	0.016	100%

4.3.6　レビューの例 2(予測値=フェイク、真値=non-フェイク)

本研究におけるフェイクレビューの特定方法 (3.2) では、フェイクであると判定されなかったレビュー (18,061) の中から、フェイクレビューの確率が高かったレビューについて確認する。

表 6　レビュー本文

No	レビュー
1	よりもはや演奏に興味があります。。。超良い密室逃脱
2	ヘビーユーザーです。お得なイベントも多いので商品を安く手に入れられるし、買い物しやすいです。 これからも使い続けます。
3	パズルができお庭も自由に変えたりできるのでとても楽しいです！
4	日本人形かわいいのこのげーむ楽しいのじゅじゅじゅ
5	パズルと庭が一緒になってて飽きない（???）?可愛い??

表 7　特徴量

No	*topic*21	*topic*37	予測確率
1	0.017	**0.05**	93.6%
2	0.017	**0.05**	70.0%
3	**0.091**	0.018	64.0%
4	**0.071**	0.018	63.4%
5	**0.071**	0.018	61.2%

表 6 において、No1 は日本語としてやや不自然な文章であり、No2 は一見すると平易なレビューに見える。共に、特徴量の重要度が 18 番目に高かったトピック 37 の値が高く、同トピックには「良い」「多い」といったポジティブな単語が含まれる。No3,4,5 はゲームのレビューであることが想定されるが、日本語として不自然であることと、表 5 と同様にトピック 21 の値が大きい。

5 まとめ

本研究では、適用が容易な方法でフェイクレビューを特定し、LDA によるトピック抽出の結果を特徴量としたランダムフォレストによるフェイクレビュー予測モデルによる検証を行った。レビュータイトルと本文の文字数に基づくモデルと比較して、LDA の結果を利用することで予測精度は大幅に向上した。フェイクレビュー予測に対して重要なトピックを特定し、予測確率に対して非線形な影響が予測力の向上に貢献することが確認できた。また、予測確率からフェイクレビューの可能性が高いレビューを発見した。

今後の課題としては、さらなる精度向上に向けレビュー本文を多角的に分析し特徴量を追加すること、フェイクレビューの特定だけでなく有益なレビューの順序付けについても検討していくことといたしたい。レビュー本文の活用方法としては、特に Android アプリにおいて、マルウェアと呼ばれるプログラムが仕込まれる問題 (ユーザーが許諾していない情報を不正に入手する、ユーザーが気付かない方法で勝手に操作を行うなど) に対応できる可能性もある。

6 謝辞

This work was supported by the Research Institute for Mathematical Sciences, an International Joint Usage/Research Center located in Kyoto University.

参考文献

[1] Griffiths, T.L. and Steyvers, M. ”Finding Scientific Topics”, in Proceedings of the National Academy of Sciences of the United States of America, 101 (Supplement 1), 5228–35,(2004).

[2] Hiroya Takamura, Takashi Inui, Manabu Okumura, ”Extracting Semantic Orientations of Words using Spin Model”, In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL2005) , pages 133–140, (2005).

[3] Steyvers M, Griffiths T. ”Probabilistic topic models” In: Landauer T, McNamara D, Dennis S, Kintsch W, editors. Latent Semantic Analysis: A Road to Meaning. Lawrence Erlbaum, (2006).

[4] 高村大也, 乾孝司, 奥村学, ”スピンモデルによる単語の感情極性抽出”, 情報処理学会論文誌ジャーナル, Vol.47 No.02 pp. 627–637, (2006).

[5] Nitin Jindal and Bing Liu. ”Review spam detection.” In Proceedings of the 16th international conference on World Wide Web (WWW ’07). ACM, New York, NY, USA, 1189-1190, (2007).

[6] Nitin Jindal and Bing Liu. ”Opinion Spam and Analysis”, Proc. International Conference on Web Search and Web Data Mining, pp.219–230 ,(2008).

[7] Cao J, Xia T, Li J, Zhang Y, Tang S,”A density-based method for adaptive LDA model selection”.(2008)

[8] 岩田具治,”潜在トピックモデルを用いたデータマイニング”, 電子情報通信学会誌,Technical Report of the 1st Workshop on Latent Dynamics,(2010)

[9] Myle Ott, Yejin Choi, Claire Cardie, Jeffrey T. Han- cock, ”Finding Deceptive Opinion Spam by Any Stretch of the Imagination”, Proceedings of the 49th Annual Meeting of the ACL, pp.309-

319,Portland,Oregon, June 19-24, (2011).

[10] Mukherjee, A., Liu, B. and Glance, N.”Spotting Fake Reviewer Groups in Consumer Reviews”, Proc. 21st In- ternational Conference on World Wide Web, pp.191– 200 ,(2012).

[11] 波部斉,”ランダムフォレスト”, 情報処理学会研究報告コンピュータビジョンとイメージメディア (CVIM),Vol.2012-CVIM-182，No.31，pp. 1-8,(2012).

[12] Sharma, K. and Lin, K.: Review spam detector with rat- ing consistency check, Proc. 51st ACM Southeast Conference, No.34 ,(2013).

[13] Igi, Makoto, Sayaka Kamei, and Satoshi Fujita , ”レビューを対象とした信頼性判断支援システムの提案.”,(2014).

[14] 中里拓哉，奥野峻弥，山名早人,”レビューサイトにおけるレビュアーの信頼性評価”, 第 6 回データ工学と情報マネジメントに関するフォーラム, (2014).

[15] 奥村学, 佐藤一誠,”トピックモデルによる統計的潜在意味解析”, コロナ社,(2015).

[16] 岩田具治,”トピックモデル”, 講談社,(2015)

[17] 三舩正暁, 金明哲 (n.d.), ”ネットショッピングにおけるスパムレビューの特徴分析”, 日本計算機統計学会第 30 回大会 p. 9-12,(2016)

[18] 松浦健太郎, ”Stan と R でベイズ統計モデリング”, 共立出版, (2016).

[19] 高島侑里 , 青野雅樹 ,”化粧品レビューサイトにおけるクチコミの有用性判定” ， 言語処理学会第 23 回年次大会発表論文集 ， pp.799-802, (2017).

[20] Greenwell, Brandon M. ”Pdp: An R Package for Constructing Partial Dependence Plots.” The R Journal 9 (1) :421–36.(2017).

[21] 佐藤敏紀, 橋本泰一, 奥村学,”単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討”, 言語処理学会第 23 回年次大会 (NLP2017).

[22] Sotaro Katsumata,Eiji Motohashi,Akihiro Nishimoto,Eiji Toyosawa(Sakihama),”Website Classification Using Latent Dirichlet Allocation and its Application for Internet Advertising”,IEEE International Conference on Data Mining Workshops, ICDMW(2017)

[23] 岡山 光平, 石川 博, 廣田 雅春, ”フェイクニュース分類器を用いた口コミサイトのレビューの分析”, 第 10 回データ工学と情報マネジメントに関するフォーラム, (2018).