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

ガイドラインに反するレビューを除去する レストラン・レビュー・サイト向けの実用的なフィルタリング処理

An Empirical Method to Remove Reviews against the Guidelines for Restaurant Review Sites

新堂 安孝 Yasutaka SHINDOH　兼村 厚範 Atsunori KANEMURA　宮尾 祐介 Yusuke MIYAO

株式会社デジタルガレージ DG Lab
DG Lab, Digital Garage, Inc.

Restaurant reviews written by customers on the Web can influence many people when they decide what to eat. Offensive or irrelevant reviews are often posted to restaurant review services and they can make people displeased and ruin services' reputation. To avoid this, restaurant review service providers issue guidelines that define what are inappropriate reviews, and employ human workers to manually remove reviews violating the guidelines. Such manual operations incur high costs and automatic filtering is desirable. Unfortunately, although several filtering methods are available, their accuracy and efficiency are still not enough to work well on actual restaurant review services because of their costs, complexities, and reviews' noisiness. In this paper, we introduce a simple, accurate, and efficient method that detects whether a review violates guidelines or not, and show through experiments on real restaurant review data that the method works well under practical and difficult situations.

1. はじめに

モバイル・インターネットの普及とともにレストラン・レビュー・サイト (e.g. Yelp[1], Retty[2]) が広く定着し，消費者が何をどこで食べるかを考える際に強い影響を持っている [1, 2]. 一方で，同サイトには後述の例[3]のような攻撃的なレビューや飲食そのものとは無関係のレビュー (以降，これらのレビューをまとめて「不適切レビュー」と呼ぶ．不適切レビュー，特に飲食そのものとは無関係のレビューは，fake review [3] に包含されないことに注意) が投稿されることが多いが，これらは，レストラン経営者などからの訴訟を招いたり，ユーザーを不快にさせサイトの品質を悪化させる要因となる．

例: 攻撃的なレビュー

注文していないお酒を不正請求されました。意図的にやっていると思います。

例: 飲食そのものとは無関係のレビュー

以前より接客が雑になっている。１４時３０分から休憩をとるのに説明されてない。

このため同サイトを運営する企業は，悪影響が出る前に不適切レビューを削除する必要があるが，その自動化が難しいため，不適切レビューを明確にすべくガイドライン (e.g. Yelp[4], Retty[5]) を用意した上で，作業者を雇用して同レビューを人手で除去せざるを得ない．この雇用に関するコストの高さは同サイトの運営の問題となる．

そこで本研究では，実レストラン・レビュー・サイトの該当コストを削減すべく，以下の 4 条件を満たす不適切レビューのフィルタリング処理を logistic regression [4] と自立語の bag-of-*n*-gram [5] (以降，それぞれ LR および BoN と表記する) を用いて開発した．

F1 同処理の実システム導入に際し，追加コストを避けるため，新しい言語リソースを必要としない．

F2 同処理を実システムに容易に導入するため，構成がシンプルである必要がある．

F3 レビュー (i.e. user-generated text) 特有の砕けた表現でも問題なく処理できる必要がある．

F4 レビュー全体に対する不適切レビューの割合 (この割合はサイトに強く依存する) が低くとも問題なく処理できる必要がある．

また，食ベログ[6] の実データを用いて評価し，同処理が実環境で高い性能を発揮することを示した．(補足: 本論文は，レストラン・レビューを主な対象に，国際会議 IEEE BigData 2018 に採択された論文 [6] を再構成したものである)

2. 関連研究

本研究に近い研究として，インターネット上の違法・有害情報 [7] を対象にしたコンテンツ・フィルタリングに関するもの [8, 9, 10] が知られている．これらの研究は，半自動生成した辞書の利用，大量の単語共起を用いた文書分類，係り受け情報を用いた文書分類などが特徴であるが，それぞれ，シード辞書は人手で生成する，語彙共起が大量になるため HDD や SSD などの追加ストレージが必要である，係り受け解析器は再学習なしでは砕けた表現で性能を発揮できない [11] という理由から，全て **F1–3** のいずれかを満たさない．また，節 1. の例「飲食そ

連絡先: 新堂 安孝 <shindo@dglab.com>

∗1 https://www.yelp.com/
∗2 https://retty.me/
∗3 本論文では食ベログの実レビューを適切に改変してレビュー例として記載する．
∗4 https://www.yelp.com/guidelines
∗5 https://retty.me/announce/tos/
∗6 https://tabelog.com/

のものとは無関係のレビュー」から容易に分かる通り，辞書を用いた手法では該当レビューの抽出が困難である．

3. 提案手法

我々は，BoN ベースの特徴ベクトルによる LR を用いてレビューが適切である (i.e. ガイドラインに反さない) 確率を算出するモデルを構築し，人手で不適切レビューを除去する作業の前に同モデルを導入する処理を提案する．同モデルが閾値以上の確率を与えたレビューは，人手の作業を介さずに適切なレビューであると判断する．

この処理により，人手の作業を大幅に減らすことが可能となる．また提案手法は，節 1. の 4 条件を以下の通り満たしている．(**F3** および **F4** については，後述の実験の結果によりその確かさが明らかになる)

F1 使用する言語リソースは既存の形態素解析辞書のみであり，新しい言語リソースを作成する必要がない．

F2 少数の比較的単純な技術 (e.g. LR, BoN) のみを用いており，構成はシンプルである．

F3 くだけた表現を入力しても，処理そのものに支障が出ない技術のみを用いている．

F4 閾値を変更する本手法は，片寄りのあるデータを適切に扱うことに向いている．

本研究では 4 種類の特徴ベクトルを比較する形で実験する．1 つのレビューに対して 1 つの特徴ベクトルを対応させ，BoN の n は「1 ($n = 1$)」または「1 と 2 の両方 ($n = 1, 2$)」を，各要素の値は「レビュー内の各 n-gram の出現頻度 (integer)」または「レビュー内に各 n-gram が存在する/しない (boolean)」を，それぞれ考える．各組み合わせの特徴ベクトルを用いた合計 4 種類のモデルを表 1 の要領で **I1**, **I2**, **B1**, **B2** と示す．

表 1: 特徴ベクトルの種類とモデル

	n-gram	
各要素の値	$n = 1$	$n = 1, 2$
integer	**I1**	**I2**
boolean	**B1**	**B2**

4. 実験

4.1 実装およびコーパス

本研究では，実験にあたり以下の要領で実装を選択した．

- 形態素解析環境: MeCab[7] 0.996, UniDic[8] 2.1.2
- LR: LIBLINEAR[9] 2.1.0

また，実験で 2 種類のコーパスを用いた．

コーパスのうち 1 つは，2005 年から 2016 年に食べログに投稿されたレビューで，適切・不適切のフラグが付与されていない．内訳は表 2 の通りである．BoN で有効な n-gram は本コーパスに含まれるものと考える．1-gram は本コーパスで 250 回以上出現した 40263 個を，2-gram は本コーパスで 100 回以上出現した 830353 個を，本研究ではそれぞれ用いた．この出現頻度による制限は，出現頻度の低い n-gram がモデルに悪影響を与えることを防ぐ狙いがある．

表 2: フラグなしコーパス内訳

レビュー数	単語数	1-gram 数
10^7	32×10^8	10^6

もう 1 つのコーパスは，2017 年 7 月から 2018 年 2 月に食べログに投稿されたレビューで，適切・不適切のフラグが人手により付与されている．内訳は表 3, 4 の通り[10]である．

表 3: フラグありコーパス内訳 (全体)

レビュー数	単語数	1-gram 数
1378273	389775791	403614

表 4: フラグありコーパス内訳 (月ごと)

収集月	レビュー数	収集月	レビュー数
2017/07	176745	2017/11	164838
2017/08	178885	2017/12	166180
2017/09	175416	2018/01	170253
2017/10	181708	2018/02	164248

4.2 実験 1

本実験では，モデル間の比較を行う．学習データとして 2017 年 7〜12 月に収集したレビューを，試験データとして 2018 年 1〜2 月に収集したレビューを，それぞれ用いた実験結果を図 1[11]に示す．

同表から分かる通り，**B2** が最も性能が良く，次いで **B1** が良い．**I2** は，**I1** よりわずかに性能が悪く，全体で見ると最も性能が悪い．曲線の形としては，**I1** と **I2** が，**B1** と **B2** が，それぞれ似ている．

B2 が高い確率 (特に true positive rate で 0.8 に該当する閾値以上の確率) を与えた false positive のレビューを見てみると，その多くは目視確認する限り食べログのガイドラインを守っている．つまり，該当レビューは true positive と考えられる．ただし以下の例の通り，かっこが非常に多い，冒頭に飲食と無関係の情報が長く続いているなど，ガイドラインを守っているか否かの判断を人手で素早く実施することが困難なものが，そのほぼ全てを占める．この傾向は，人手の適切・不適切の判断にばらつきがあること，同時に **B2** が学習データのノイズに対して頑強であることを，それぞれ意味する．また **B2** は節 1. の条件 **F3**, **4** を十分に満たしていると考えられる．

∗7 http://taku910.github.io/mecab/
∗8 https://unidic.ninjal.ac.jp/
∗9 https://www.csie.ntu.edu.tw/~cjlin/liblinear/
∗10 適切・不適切の内訳は食べログの機密情報のため掲載不可．
∗11 本研究では全ての実験結果を ROC 曲線を用いて示す．

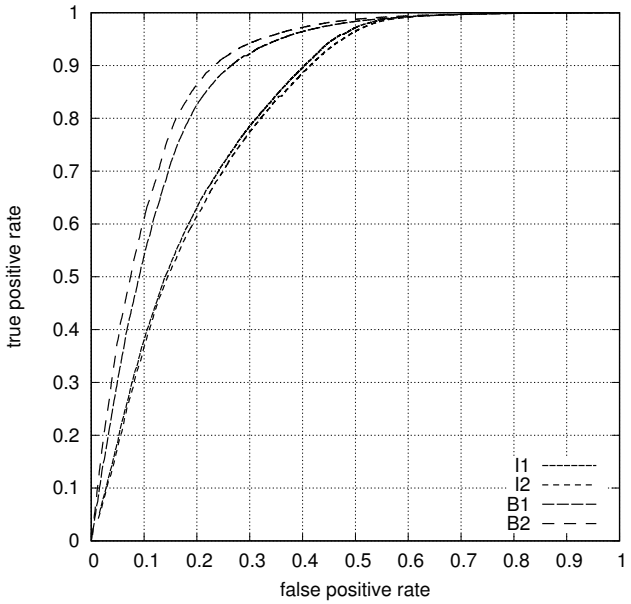

図 1: 実験 1 (各モデルの出力)

例: かっこが非常に多いレビュー

賽（さいころ）に容子（やうす）似（に）せたる “黒砂糖（くろざとう）”
無農薬栽培（あやしげなるくすりぬき）の “紅茶（こふちや）”
肥後天草（ひごあまくさ）の “春子（かすご）”・“小鰭（こはだ）”

例: 冒頭に飲食と無関係の情報が続いているレビュー

福島県会津若松市出身の歴史上の要人は数多い。それは今でもよく知られているが、その偉業は忘れられている。
［*skip over many sentences similar to the above sentences*］
今回食べたソースかつ丼はこの店の名物だ。本当に上質の豚肉を使っていた。
［*skip over many sentences that mention foods*］
とてもおいしかった。再訪したい。

4.3 実験 2

本実験では，各モデルが学習データの量に対してどうふるまうかを確認する．各学習データ (以降，1mo, 2mo, 4mo, 6mo と示す) として表 5 の期間に収集したレビューを，試験データとして 2018 年 1〜2 月に収集したレビューを，それぞれ用いた．各モデルの実験結果を図 2–5 に示す．

表 5: 学習データ (実験 2)

収集期間	1mo	2mo	4mo	6mo
開始月	2017/12	2017/11	2017/09	2017/07
終了月	2017/12	同左	同左	同左

I2 (図 3) のみ，学習データが増えれば増えるほど性能が悪化している．これは，2-gram の総数が大きいことと各 *n*-gram の出現頻度がまちまちであることから，**I2** が複雑になりすぎた結果，古いレビューに強く適合する悪影響が発生した可能性が考えられる．また，食べログのレビューが含む季節に強く依存した表現 (e.g. 蒸し暑い，冷やしそうめん) も，悪影響を与え

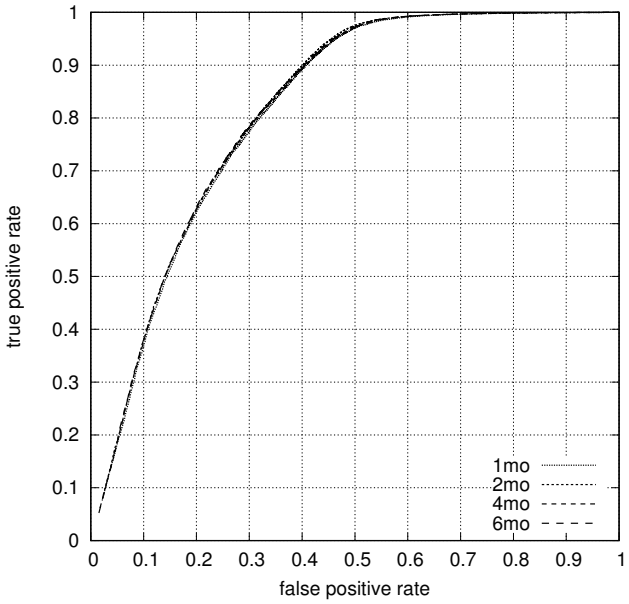

図 2: 実験 2 (I1 の出力)

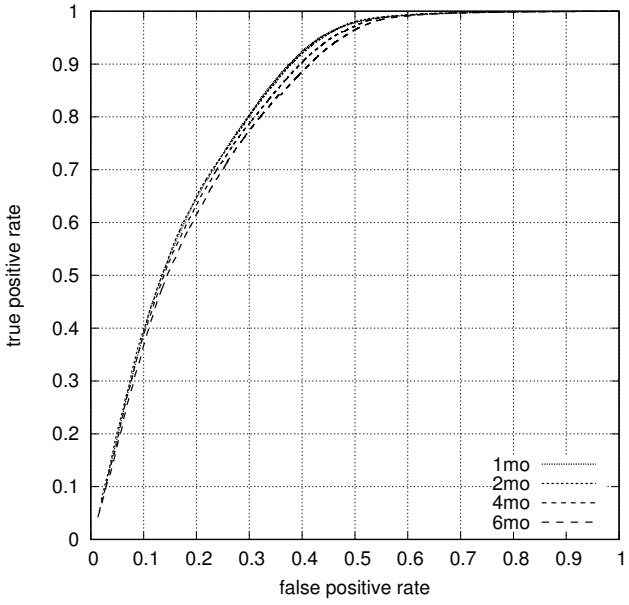

図 3: 実験 2 (I2 の出力)

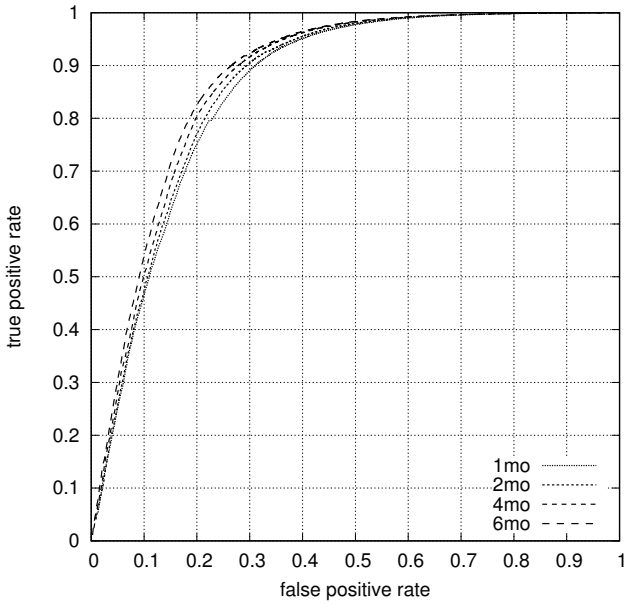

図 4: 実験 2 (B1 の出力)

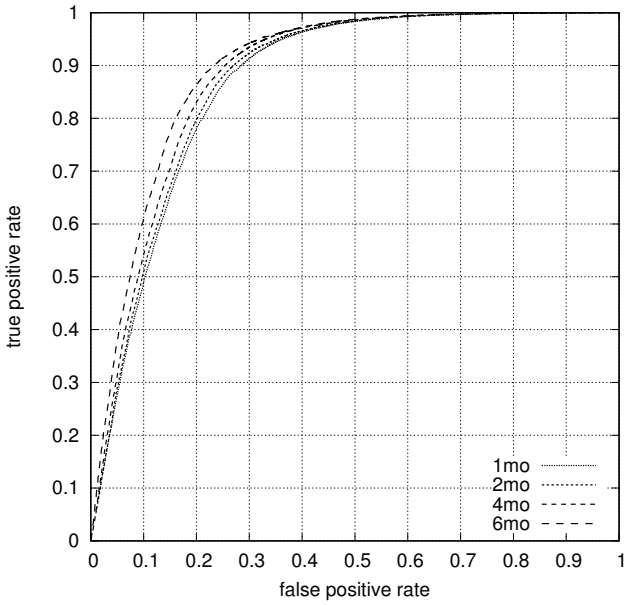

図 5: 実験 2 (B2 の出力)

た可能性がある．

一方で，**B1** (図 4) および **B2** (図 5) は，学習データが増えれば増えるほど性能が向上している．これは **B1** と **B2** の汎化性能が高いことを示していると考えられる．また **B1** と **B2** は，1ヶ月分の学習データ (1mo) を用いた結果でも，メール・フィルタリングで用いられる手法に近い **I1** より良い性能が出ている．

5. まとめ

本研究では，レストラン・レビュー・サイトに投稿される不適切レビューを人手で除去するコストを削減すべく，該当レビューを除去する実用的な手法を提案し，食べログの実レビュー・データを用いて実験を行い以下の通りその有用性を示した．

- 実験を通し，不適切レビューを人手で除去するコストの 80%程度を，最良のモデル **B2** で削減できる[12]ことが分かった．また同モデルについては，学習データが増えることで性能が向上する傾向を持つことが分かった．(節 4.2 および節 4.3 を参照のこと)
- 実システムに組み込む際に問題となる要素を考慮することで，実システム導入が容易な設計を実現できた．(節 1. および節 3. を参照のこと)

参考文献

[1] MyVoice Communications, Inc. ネット上の口コミ情報に関するアンケート調査 (第 4 回). https://myel.myvoice.jp/products/detail.php?product_id=22515, 2017.

[2] Japan Finance Corporation. 外食に関する消費者意識と飲食店の経営実態調査. https://www.jfc.go.jp/n/findings/pdf/seikatsu25_1218a.pdf, 2013.

[3] Arjun Mukherjee, Bing Liu, and Natalie Glance. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on World Wide Web*, 2012.

[4] David Cox. The regression analysis of binary sequences (with discussion). *Journal of the Royal Statistical Society B*, Vol. 20, pp. 215–242, 1958.

[5] Alice Zheng and Amanda Casari. *Feature Engineering for Machine Learning*. O'Reilly Media, Inc., 2018.

[6] Yasutaka Shindoh, Atsunori Kanemura, and Yusuke Miyao. A simple method to remove reviews against guideline for online review services. In *Proceedings of the 2018 IEEE International Conference on Big Data*, 2018.

[7] Ministry of Internal Affairs and Communications. インターネット上の違法・有害情報に対する対応. http://www.soumu.go.jp/main_sosiki/joho_tsusin/d_syohi/ihoyugai.html, 2009.

[8] Tatsuya Ishisaka and Kazuhide Yamamoto. Detecting nasty comments from BBS posts. In *Proceedings of The 24th Pacific Asia Conference on Language, Information and Computation*, pp. 645–652, 2010.

[9] Kenji Nakamura, Shigenori Tanaka, Yuhei Yamamoto, and Satoshi Abiko. Method of filtering harmful information considering extraction range of word co-occurrence (in Japanese). *IPSJ Journal*, Vol. 54, No. 2, pp. 571–584, February 2013.

[10] Kazushi Ikeda, Tadashi Yanagihara, Gen Hattori, Kazunori Matsumoto, and Yasuhiro Takishima. Hazardous document detection based on dependency relations and thesaurus. In *AI 2010: Advances in Artificial Intelligence*, pp. 455–465, 2010.

[11] Satoshi Namba, Kenta Kadouchi, Yasuhiro Tajima, and Genichiro Kikui. マイクロブログに対する文境界推定および係り受け解析. In *Proceedings of the 21th Annual Meeting of the Association for Natural Language Processing*, 2015.

∗12 ここでは不適切レビューの総数が適切レビューの総数より十分に小さいと仮定.