評価付き自然言語データの定量分析

——どのような消費者レビューが「参考になった」を集めるのか?——

金 勝 鎮, 多 田 伶, 勝 又 壮太郎

A Quantitative Model for Natural Language Dataset with Evaluation Information:
What Topic makes Consumer Reviews “Helpful”?

Seungjin KIM, Rei TADA and Sotaro KATSUMTA

This study proposes LDA with evaluation information. The authors focus on online consumer reviews with “helpfulness of review”. The purpose is to find specific topics that affect consumer’s decision making from many consumer reviews. The authors exploit an evaluation score in addition to review text which is described in the natural language for simultaneously classifying topic model. The “helpfulness” scores are accessed by readers and are with integers of zero or more. The authors apply the suggested model in reviews for coffeemakers crawled from consumer’s review website. The results show that the model distinguishes two topic groups with extremity which indicates that is helpful or not. Also, it shows that the two topic groups also have qualitative differences among firms (brands). The authors discuss the application of the model and findings for both review website managers and firm’s marketers.

Key words: Topic modeling, Consumer reviews with evaluation, Unstructured data
キーワード: トピックモデル, 評価付き消費者レビュー, 非構造データ

1. は じ め に

消費者行動を定量的に解析し, 説明・予測するためには, 購買行動や消費金額などの定量的なデータだけでなく, 購買に至る感情を示した SNS の投稿や, 購買後の満足度などが表明されたレビューデータなどの「非構造データ (unstructured data)」を検討していく必要がある (Balducci & Marinova, 2018). とくに, インターネット上の消費者発信データが幅広く入手可能になった近年, 多様な非構造データをどうやって企業経営に生かしていくかが大きな課題となっている (Lamberton & Stephen, 2016). 質的データ分析においては, ある程度科学的に定量化するための確立され

た方法もあるが (たとえば, Krippendorff, 2012), 定量化のプロセスにおいて分析者が目視でデータを確認する必要もあり, 大規模なデータの分析には向かない. また, インターネット上には, 長期間にわたり大量の消費者が発信した自然言語データが蓄積されており, 分析者が個別のデータを見て判断することは困難であることも多い. そこで, 質的なデータを量的なデータに変換し, 定量分析を行う手法が求められる (Humphreys & Wang, 2018; Wedel & Kannan, 2016). たとえば, 本研究の目的である自然言語データの分析においては, 意味を持つ最小単位である形態素でテキストデータを分解することで, 自然言語データの定量分析が可能になる. 近年, マーケティング分野において非構造データの分析方法に対する議論が進展しており, 非構造データを定量化して分析を行う方法が広く提案されている (Berger, Humphreys, Ludwig, Moe, Netzer, & Schweidel, 2019). とくに, 自然言語データの分析においては, LDA (Latent Dirichlet Allocation; Blei, Ng, & Jordan, 2003) をはじめとしたトピックモデ

ルが注目されており、近年多くの応用研究が見られている (Tirunillai & Tellis, 2014; Büshkin & Allenby, 2016; Liu, Burns, & Hou, 2017).

LDA は、複数の文書の分類を行うための統計モデルであり、文書中出现する語彙の共通性に注目して文書をいくつかのトピックに分類するモデルである。文書中の単語の出現確率にカテゴリカル分布を仮定し、その事前分布として共役分布であるディリクレ分布を仮定している。LDA の実用上の大きな利点は推定方法にあり、Griffiths & Steyvers (2004) によって提案された Collapsed Gibbs sampling を使うことができる。

しかしながら、インターネット上で得られるデータについては、自然言語データだけでなく、様々な定量データも合わせて検討する必要がある。たとえば、ある消費者の発信したレビューの質 (Filieri, 2015) を測る指標となりえる「いいね」数やレビューが「参考になった」と評価された回数 (以下、「参考数」) などを考慮して定量的な分析を行うことで、どのようなトピックが閲覧者の興味を引き、必要となる情報であるかを検討することができる。同じレビューサイトの自然言語データであっても、その質が大きく異なる可能性がある。消費者が意思決定をより容易にするため、レビューを利用する (Dabholkar, 2006) とともに、Jiang & Benbasat (2004) が主張したように、「点数」、「いいね」またはレビューの「参考数」などのレビューに付加的に与えられる評価情報は消費者の購買意思決定において製品評価に影響を与える。たとえば、Poston & Speier (2005) は情報処理と情報検索の観点から、「点数」がレビューの内容を把握するときに有用な情報として意味があることを明らかにした。Eslami, Ghasemaghaei, & Hassanein (2018) と Filieri (2015) も個々のレビューの評価点数と総合評価点数が消費者に影響を与えることを示した。López-López & Parra (2016) は「参考数」が高いレビューの評価点数と製品の総合評価点数の一致性によって他の消費者に与える影響が違うことを明らかにし、「参考数」が、消費者が製品を評価する重要な基準として認識されていることを示した。また、「参考数」に影響を与える要因として、レビューの投稿者 (Racherla & Friske, 2012), レビューの文書特性 (Eslami et al., 2018; Mudambi & Schuff, 2010; Racherla & Friske, 2012; Pan & Zhang, 2011), 製品の特性 (Mudambi & Schuff, 2010) が議論されてきた。Chen & Lurie

(2013) はレビューの内容の中で時間の表現に注目し、レビュー内の時間を表す表現 (たとえば、「昨日」、「買ってから 3 カ月」など) が「参考数」にポジティブな影響を与えることを明らかにした。しかし、彼らの研究はある特定の単語に着目したため、レビュー全体の動向または内容と「参考数」の関係を具体的に把握することは難しい。

そこで、本研究では、文書内で観測される語彙に基づいてトピックへの所属確率を推定する LDA を拡張して、文書のレビューに付されている「参考数」も反映してトピックへの所属確率を推定するトピックモデルを提案する。また、レビューの自然言語データから「参考数」を予測する方法についても議論する。そのため、レビューサイトから収集したテキストデータにおいては形態素分析を行い、打消し表現や指示代名詞、1 文字の単語のような単独では意味を持たない単語を除外するなどの前処理を行うことで定量分析のためのデータセットを用意する。そして、テキストデータと同時に収集したレビューごとの「参考数」を用いて提案トピックモデルの検討や推定を行う。また、レビューの自然言語データから「参考数」を予測する方法についても議論する。最後に、提案モデルから得られた結果について議論する。

2. モデル

2.1. データの表現

本節ではモデルの概要を示す。以下では、Blei et al. (2003) によって提案された LDA および、これを拡張した Katsumata & Kim (2020) をもとに、「参考数」の変数を組み込んだモデルを定義する。

まず、レビューデータのノテーションを定義する。レビューデータは D 個存在し、各レビューを「文書 (document)」と呼ぶ。文書 d には自然言語で書かれたテキストと、そのテキストの情動的価値を他者が評価した「参考数」が付されている。テキストについて、データセットから得られるすべての「単語 (word)」の総数は N とおく。ここで、第 i 番目 ($i = 1, \dots, N$) で観測される単語を w_i とおく。ここで、単語は V 種類ある「語彙 (vocabulary)」のいずれか 1 つが観測されるとする。したがって、 w_i は V 次元のベクトルとなる。第 i 番目の単語 w_i の要素 w_{iv} について、語彙 v が観測されたとき、 $w_{iv} = 1$ となり、 $v' \neq v$ のとき $w_{iv'} = 0$ となる変数として定義する。また、デー

タセットに登場する全ての単語 w は D 個あるいずれか 1 つの文書に所属することになる．この単語と文書の所属関係を示す変数として x_i を導入する． x_i は D 次元のベクトルであり，第 i 番目の単語が文書 d に所属していたときには $x_{id} = 1$ をとり， $d' \neq d$ について $x_{id'} = 0$ をとる変数として定義できる．また，文書 d に出現する単語の数を N_d とおく． $N_d = \sum_{i=1}^N x_{id}$ である．文書ごとに与えられている評価について，文書 d に付されている「参考数」を y_d とおく．「参考数」は 0 以上の整数である．

2.2. LDA：文書のトピック分解

LDA は，単語 w_i が以下のような V 次元のカテゴリカル分布に従っていると仮定するモデルである．

$$w_i \sim \text{Categorical}_V(\tilde{\phi}_i) \quad (1)$$

パラメータ $\tilde{\phi}_i$ については潜在的に K 種類のトピックが存在すると仮定する．モデルは，文書が K 種類の話題に分類されると仮定しており，それぞれの話題は一般に「トピック」と呼ばれている（たとえば，岩田 2015）．このとき，単語 i が所属するトピックを示す潜在変数を z_i とおく． z_i は K 次元のベクトルであり第 i 番目の単語がトピック k に所属しているときは $z_{ik} = 1$ をとり， $k' \neq k$ のとき， $z_{ik'} = 0$ となる．これを組み合わせてパラメータ $\tilde{\phi}_{iv}$ の要素を以下のように定義する．

$$\tilde{\phi}_{iv} = \prod_{k=1}^K \phi_{kv}^{z_{ik}} \quad (2)$$

$\tilde{\phi}_{iv}$ は単語 i の出現確率のパラメータで， K 個のトピック毎から語彙の選択確率 $\phi_{kv}^{z_{ik}}$ の同時確率で定義する．ここで， ϕ_{kv} はトピック k から語彙 v の生成確率である．

また， z_{ik} については，以下のような K 次元のカテゴリカル分布に従っていると仮定する．パラメータ $\tilde{\theta}_i$ はトピックの生成確率であり， $\tilde{\theta}_{dk}$ は文書 d にトピック k が割り当てられる確率である．

ここで， x_i は上で定義した文書の所属を示す変数である．

$$z_i \sim \text{Categorical}_K(\tilde{\theta}_i), \tilde{\theta}_{ik} = \prod_{d=1}^D \theta_{dk}^{x_{id}} \quad (3)$$

事前分布のパラメータ ϕ_k および θ_d については，カテゴリカル分布と共役であるディリクレ分布に

従っていると仮定する． $\phi_k \sim \text{Dirichlet}(\beta)$ ， $\theta_d \sim \text{Dirichlet}(\alpha)$ とする．ここで， β と α は各 ϕ_k と θ_d のパラメータである．

2.3. 「参考数」の仮定

続いて，LDA に加えて，「参考数」を同時に分析するためのモデル拡張について説明する．文書 d には「参考数」が示されている．ここで，文書 d が得られた「参考数」を y_d とし， y_d は 0 以上の整数値をとる変数なので，幾何分布を仮定する．幾何分布は離散確率分布であり，0 から無限大までの値の確率を表現できる．また，幾何分布は共役事前分布がベータ分布であり，カテゴリカル分布とディリクレ分布を仮定する LDA と共通した性質をもつため，推定が容易になる利点がある．

$$y_d \sim \text{Geometric}(\tilde{\rho}_d) \quad (4)$$

また，「参考数」を規定するパラメータ $\tilde{\rho}_d$ も単語のパラメータ $\tilde{\phi}_i$ と同様に潜在的に存在する K 種類のトピックに依存して値が異なると仮定する．文書 d に付される「参考数」 y_d が所属するトピックを示す潜在変数を導入して，これを c_d とすると，単語の所属するトピックを示す潜在変数 z_i と同様に， $\tilde{\rho}_d$ を以下のように定義することができる．

$$\tilde{\rho}_d = \prod_{k=1}^K \rho_k^{c_{dk}} \quad (5)$$

$c = \{c_{dk}\}$ ， $\rho = \{\rho_k\}$ として，「参考数」 y_d の質量関数を以下のように定義する．

$$\pi(y_d | \rho, c) = \tilde{\rho}_d^{y_d} (1 - \tilde{\rho}_d) \quad (6)$$

ここで， y_d は $(1 - \tilde{\rho}_d)$ ではなく $\tilde{\rho}_d$ の冪指数とすることで， y_d の期待値と $\tilde{\rho}_d$ が正の関係となるようにする．パラメータ $\tilde{\rho}_k$ は $(0, 1)$ の値をとり， y_d の期待値は $\tilde{\rho}_d / (1 - \tilde{\rho}_d)$ となる．したがって， $\tilde{\rho}_d$ が大きいほど「参考数」が多くなるという構造になっている．

また， c_d については z_i と共通のパラメータ θ_d を持つとして，以下のようにカテゴリカル分布を定義する．

$$c_d \sim \text{Categorical}_K(\theta_d) \quad (7)$$

パラメータ ρ_k については，幾何分布と共役なベータ分布に従うと仮定する． $\rho_k \sim \text{Beta}(\gamma)$ とおく．ここで， γ は事前分布 ρ_k のパラメータである．

2.4. 推定

完全条件付き事後分布は以下ようになる.

$$\begin{aligned}\pi(z, c, \phi, \theta, \rho | w, y) &\propto \pi(w | z, \phi) \pi(z | \theta) \pi(\phi) \\ &\quad \times \pi(y | c, \rho) \pi(c | \theta) \pi(\rho) \\ &\quad \times \pi(\theta)\end{aligned}\quad (8)$$

この完全条件付き事後分布から, パラメータである ϕ , θ , ρ を積分によって消去し, z と c についての閉じた形の事後分布を得ることができる. これを Collapsed Gibbs sampling とよぶ. Griffiths & Stayvers (2004) によると, Collapsed Gibbs sampling は高速な推定が可能であり, 収束も早い.

$$\pi(z, c | w, y) \propto \int \int \int \pi(z, c, \phi, \theta, \rho | w, y) d\phi d\theta d\rho \quad (9)$$

これをもとに z_i の事後分布を導出していくと, $z_{ik} = 1$ をとる確率は以下の式から得ることができる.

$$\pi(z_{ik} | z_{-ik}, c, w, y) \propto \frac{n_{kv^*, -i} + \beta}{n_{k, -i} + V\beta} (n_{d^*k, -i} + c_{d^*k} + \alpha) \quad (10)$$

ここで, v^* は第 i 番目に観測された単語であり, $w_{iv^*} = 1$ である. また, d^* は第 i 番目の単語が所属する文書であり $x_{id^*} = 1$ である. また, $n_{kv^*, -i} = \sum_{i'=1, i' \neq i}^N w_{i'v^*} z_{i'k}$, $n_{k, -i} = \sum_{v=1}^V n_{kv, -i}$, $n_{d^*k, -i} = \sum_{i'=1, i' \neq i}^N x_{i'd^*} z_{i'k}$ である. ここで, $-i$ はデータセットのすべての単語 w から単語 w_i を除外することを示す.

また, これをもとに c_d の事後分布を導出していくと, $c_{dk} = 1$ をとる確率は以下の式から得ることができる.

$$\begin{aligned}\pi(c_{dk} | c_{-dk}, z, w, y) &\propto \frac{\prod_{q=1}^{y_d} (m_{k1, -d} + q - 1 + \gamma) (m_{k0, -d} + \gamma)}{\prod_{q=1}^{y_d+1} (m_{k, -d} + q - 1 + 2\gamma)} \\ &\quad \times \prod_{t=1}^{n_{dk}} (t + \alpha)\end{aligned}\quad (11)$$

ここで, $m_{k1, -d} = \sum_{d'=1, d' \neq d}^D y_{d'} c_{d'k}$, $m_{k0, -d} = \sum_{d'=1, d' \neq d}^D c_{d'k}$, $m_{k, -d} = m_{k1, -d} + m_{k0, -d}$ である. また, $-d$ は $-i$ と同様に, データセットのすべての文書 D から文書 d を除外することを示す.

上記の事後分布から z と c が得られていれば, ここからパラメータ ϕ_k , ρ_k , θ_d を以下のように得ること

ができる.

$$\phi_{kv} = \frac{n_{kv} + \beta}{n_k + V\beta} \quad (12)$$

$$\rho_k = \frac{m_{k1} + \gamma}{m_k + 2\gamma} \quad (13)$$

$$\theta_{dk} = \frac{n_{dk} + c_{dk} + \alpha}{N_d + 1 + K\alpha} \quad (14)$$

ここで, $n_{kv} = \sum_{i=1}^N w_{iv} z_{ik}$, $n_k = \sum_{v=1}^V n_{kv}$, $m_{k1} = \sum_{d=1}^D y_d c_{dk}$, $m_{k0} = \sum_{d=1}^D c_{dk}$, $m_k = m_{k1} + m_{k0}$, $n_{dk} = \sum_{i=1}^N x_{id} z_{ik}$ である. n_{kv} はトピック k に属する単語 v の数, n_k はトピック k に属する単語の数を示す. m_{k1} はトピック k に属する文書の y_d の合計, m_{k0} はトピック k に属する文書の数である. n_{dk} は文書 d の単語のうちトピック k に属する単語数を示す.

2.5. 予測

加えて, 予測について説明する. 「参考数」の文書評価 y を文書の単語 w から説明することもできる. 欠損があれば, 以下の分布に従って乱数を発生させることで推定値を得ることができる.

$$y_d | c_d \sim \text{Geometric}(\tilde{\rho}_d) \quad (15)$$

$$\tilde{\rho}_d = \prod_{k=1}^K \left(\frac{m_{k1, -d} + \gamma}{m_{k, -d} + 2\gamma} \right)^{c_{dk}} \quad (16)$$

幾何分布に従う乱数については, 同様のパラメータ $\tilde{\rho}$ に従うベルヌーイ分布から乱数を発生させ, 1 が出るまで繰り返し, 0 の出た回数を y_d の乱数による生成値とすればよい.

本研究では, 単語 w と文書評価 y の2つをインプットデータとする LDA の拡張モデルを文書評価付き LDA とよぶ. グラフィカルモデルによる LDA および文書評価付き LDA の違いは図 1 の通りであり, 図に用いられる記号の説明は表 1 の通りである. 文書数は D , 単語数の合計は N であるが, 「参考数」の所属トピックを規定する c および出現単語の所属トピックを規定する z が同じ θ を共有している. このように複数のインプットデータを仮定した研究として, たとえば Mimno, Wallach, Naradowsky, Smith, & McCallum (2009) がある. 共役分布を仮定しているため, Collapsed Gibbs sampling によるパラメータ推定が可能になっている.

LDA

文書評価付きLDA

図 1. モデル概念図

表 1. 記号の説明

記号	説明
D	文書数
N_d	文書 d に含まれる単語数
K	トピック数およびトピック数
θ_{dk}	文書 d でトピック k が割り当てられるパラメータ
ϕ_{kv}	トピック k から語彙 v が生成されるパラメータ
ρ_k	トピック k から「参考数」 y が生成されるパラメータ
w_i	文書 d の単語 i
z_i	単語 i に割り当てるトピックの潜在変数
c_d	文書 d の「参考数」に基づいて割り当てるトピックの潜在変数
α, β	ディリクレ分布のパラメータ
γ	ベータ分布のパラメータ

図 2. 「参考数」のヒストグラム

3. 実 証 分 析

3.1. データ概要

実証分析に用いるデータは kakaku.com から収集したユーザーレビューである。ユーザーレビューには、自然言語で記述されたレビュー内容と、そのレビューを読んだ閲覧者が付ける「参考数」が記録されている。kakaku.com では様々なカテゴリーの製品レビューを閲覧することができるが、本研究では、コーヒーメーカーを対象として分析を行う。データは 2019 年 11 月 25 日に収集しており、合計 3095 件のユーザーレビューが得られた。最新のレビューは 2019 年 11 月 25 日に投稿されている。また、最も古いレビューは 2005 年 12 月 7 日に投稿されている。ここで、レビューに本文のないものを除去して 2969 件を分析対象とする。したがって、 $D = 2969$ である。この 2969 件にはすべて「参考数」が 0 以上の整数値で得られて

いる。ここで、「参考数」が 0 であったレビューの数は 897 件 (30.2%) であり、約 7 割のレビューには 1 人以上の閲覧者から「参考になった」がチェックされている。2969 件の「参考数」の概要について、中央値は 1、算術平均値は 2.73、最大値は 60 であり、ヒストグラム (図 2) はほぼ単調減少であり、幾何分布を仮定することが適切といえる。

3.2. 自然言語データの定量化

「参考数」はそのまま定量データとして分析することができるが、自然言語データは定量化しなければならない。本研究では、定量化においては、Kudo, Yamamoto, & Matsumoto (2004) によって開発された MeCab を使って品詞分解を行い、名詞、動詞、形容詞のみを抽出した。ただし、単語単体では打消し表

現を検証することができないので、打消しを検知する処理を追加する。品詞分解を行って、単語の列を得た後、名詞および形容詞の後に打消し表現の助詞を検知した場合、直前の単語に「(-1)」を付加し、打消しであることを明示する。たとえば「使えない」という表現は「使う + ない」と分解できるが、「ない」が打消し表現であるため、「使う (-1)」という単語が分析に用いられる。これは「使う」とは異なる単語として処理する。

次に、1文字の単語、1回しか観測されていない単語、「この」や「もの」などの単体では意味を持たないが多く用いられている指示代名詞などを除外する。得られた語彙数は $V = 6685$ であった。また、単語数の合計については、 $N = 179092$ となった。ここで、単語数の最も少ない文書は3単語、最も多い文書は474単語、算術平均は59.98単語、中央値は46単語であった。

3.3. 予測とモデル比較

LDAは文書を複数のトピックに分類するモデルであり、本研究で用いる文書評価付きLDAも同様の構造をもつ。したがって、最適なトピック数を決定するために、異なるトピック数を仮定したいくつかのモデルを並行して推定し、結果を比較する必要がある。本研究では、 $k \in \{1, 2, 5, 10, 20, 30, 40, 50, 60, 80, 100, 150\}$ の合計12個のモデルを比較対象とする。

モデルの比較について、Griffiths & Steyvers (2004) では対数周辺尤度 (サンプリングした尤度の調和平均) が用いられており、本研究でもこれを推定して考慮する。しかしながら、LDAにおいて対数周辺尤度を基準にすると、トピック数 K は大きくなる傾向にあり (たとえば、Griffiths & Steyvers (2004) ではトピック数が300で対数周辺尤度が一番高くなっている)、分析者が得られたトピックを考察するには多すぎる場合もある。また、予測モデルとして考えた場合、対数周辺尤度だけでなく予測能力を比較する必要がある。そこで、本研究では、2014年以降に投稿されたレビューの一部を予測用セットとして、「参考数」を欠損扱いで推定し、予測精度を測定する。2014年以降、毎年のレビューの半数について、単語 w は得られるが「参考数」 y を得ることができない状態にして、 z 、 c に加えて欠損となっている y も推定していく。予測用のレビューは856件であり、レビュー全体の28.8%である。

推定において、イタレーション回数は合計2000回で、そのうちの前半1000回を稼働検査期間として棄て、後半の1000回をサンプルとして取得する。また、事前パラメータについて $\alpha = 50/K$ 、 $\beta = 0.1$ 、 $\gamma = 0.1$ とおいた。 z 、 c の初期値はランダムに割り当てられる。

4. 分析結果

4.1. モデル比較結果

まずは、トピック数 K を決めるためにモデルの比較を行う。モデルの比較にはAUCを用いる。AUCは予測能力を検討する指標として広く使われている。AUCは0から1の間の値を取り、完全にランダムで予測を行ったときの期待値が0.5であり、0.5よりも大きければ予測能力があるといえる絶対的な基準を持っているので、指標の解釈が容易である。

AUCはROC曲線の曲線下面積であり、値が高いほど予測能力が優れていることを示している。また、ROC曲線は $\{0, 1\}$ の2値の判別問題の予測精度を比較するための指標なので、本研究のトピック数 K を決めるためのROC分析では「参考数」が1件以上あれば1、そうでなければ0として比較を行っている。比較のための指標としては $\bar{\rho}$ の推定値を用いている。また、予測のための得点については、毎回のイタレーションで ρ_k と c_d が得られるため、第 (h) 回目のイタレーションによって得られる $\bar{\rho}^{(h)} = \prod_{k=1}^K (\rho_k^{(h)})^{c_{dk}^{(h)}}$ としてサンプルを集め、その事後平均値を用いている。幾何分布のパラメータ ρ と期待値は単調な関係にあるため、パラメータが得られればROC曲線を描くことができる。また、図中の黒い点は予測用データに対する結果であり、アウトオブサンプルフィットを示している。また、図中の白い点は y が得られているデータに対する結果であり、インサンプルフィットを示している。

図3は得られた対数周辺尤度と予測結果の比較である。左が対数周辺尤度であり、右がROC曲線の曲線下面積でAUCによる結果である。まず、対数周辺尤度の結果からは、 $K = 50$ のときに最も当てはまりがよいという結果が得られた。

AUCの結果をみると、インサンプルフィットはトピック数 K の増加とともに単調に増加しているため、あまりに K が大きいとオーバーフィッティングをし

図 3. モデル比較結果

ている可能性があり、比較に用いることは適切でないといえる。アウトオブサンプルフィットの結果を見ると、トピック数 $K = 1$ のときは 0.5 を下回っているが、 $K = 2, 5, 10, 20$ までは単調に増加していることがわかる。しかしながら、 $K = 30$ では $K = 20$ よりも予測精度が低く、その後も大きな改善はないようである。

対数周辺尤度からは $K = 50$ が最も適合度が高いという結果が得られたが、AUC のアウトオブサンプルフィットを見ると $K = 20$ が最も予測精度が高いという結果が得られた。 $K = 20$ 前後の対数周辺尤度を確認すると、 $K = 20$ までは急激に値は改善していくが、 $K = 20$ を超えると比較的対数周辺尤度の改善が緩やかになっていることがわかる。したがって、本研究では、最も予測精度が高く、ある程度対数周辺尤度も改善しており、分析者による解釈が容易であるという観点から、 $K = 20$ のモデルに注目して考察していく。

4.2. 結果の考察

表 2 は、 $K = 20$ のモデルから得られたトピックについて、「参考数」の集めやすさのパラメータである ρ_k および、関連の強い単語 (ϕ_k の各要素 ϕ_{kv} について、大きいもの上位 10 単語) を掲出したものである。上から ρ_k の一番大きなトピックをならべており、最下段が最も ρ_k の小さいトピックとなっている。したがって、上にあるほど相対的に「参考数」を集めやすく、下にあるほど「参考数」を相対的に集めにくいトピックである。ここで、推定された 20 個のパラメー

タ ρ_k のトピックは「参考数」だけではなく、文書内の単語のトピックにも連動しているので、 ρ_k とトピックは意味的に「トピックの相対的な傾向」を示す。

表 2 を見るとわかるが、相対的に最も「参考数」を集めやすいトピックは Topic No. 4 であり、故障や交換に関する内容を記述したレビューである。閲覧者が故障対応について情報を求めていることがわかる。また、次に「参考数」を集めやすいトピックは、「掃除」や「面倒」などの単語が頻出する Topic No. 17 であった。このトピックについては、継続的に使用した内容が記載されているレビューといえる。これらのトピックに共通していることは、実際にある程度の期間当該製品を使用し、故障や交換を経験したり、毎回の作業に面倒を感じたりするなど、継続的な使用者でないといけない点を記載しているところであるといえる。コーヒーマーカーも耐久消費財の一つであり、購入してから継続的に利用することで価値が評価される製品である。また、通常はメーカーの発信する情報や広告などでは交換や日々のルーチンとなる作業を前面に出すことは少なく、この面でも他のメディアから得られない情報が「参考数」を集めやすい要因であると考えられる。

一方、下段のトピックを見ると、相対的に最も「参考数」を集めにくい Topic No. 11 については「簡単」、「満足」、「手軽」など、好意的な単語が多く表れているが、メーカーが発信する情報と大きな差はないと閲覧者が評価している可能性がある。また、Topic No. 6 についても、デザインに関する単語やサイズに関する

表 2. トピック別の ρ と頻出単語

Topic No.	ρ	頻出単語
4	0.769	使用, メーカー, ところ, 壊れる, 故障, 悪い, 交換, くる, 修理, 今回
17	0.762	フィルター, ミル, 挽く, コーヒーメーカー, 全自動, ペーパー, 付き, 掃除, 毎回, 面倒
5	0.748	ドリップ, 時間, お湯, かかる, 必要, ハンド, ため, 手間, 蒸らす, 落ちる
10	0.694	カプセル, 楽しめる, 種類, 一杯, カフェ, 比べる, 簡単, 飲む, 1 杯, ドルチェグスト
13	0.683	タンク, 部分, 本体, にくい, 給水, 外す, スペース, 洗う, 部品, 構造
12	0.671	飲む, お湯, インスタント, インスタントコーヒー, パリスタ, 違う, カフェラテ, 手間, ブラック, 牛乳
19	0.667	機種, 洗浄, 設定, 捨てる, 抽出, 全自動, カス, 掃除, デロンギ, モデル
3	0.657	保温, ポット, サーバー, ステンレス, ガラス, コーヒーメーカー, 注ぐ, タイプ, 洗う, 容器
20	0.655	れる, 抽出, カップ, 感じる, 温度, しれる (-1), 程度, 入る, 場合, 書く
15	0.654	コーヒー, 美味しい, 飲む, られる, 香り, 珈琲, 好き, 楽しむ, れる, お勧め
2	0.652	エスプレッソ, ミルク, マシン, カプチャーノ, 作れる, 操作, 必要, ボタン, 面倒, スチーム
18	0.629	ない, 機能, 製品, 評価, ため, 考える, 以上, 問題, 非常, ところ
1	0.619	入れる, ボタン, 電源, 押す, セット, スイッチ, 便利, おく, 自動, あと
7	0.580	使う, 買う, コーヒーメーカー, みる, 安い, てる, 作る, 見る, 言う, そう
9	0.564	できる, しまう, 自分, とき, 濃い, 好み, できる (-1), 出る, くる, 調整
16	0.558	良い, 思う, 出来る, 無い, 大きい, 欲しい, 付く, 残念, 使い勝手, 出来る (-1)
8	0.549	購入, 商品, 満足, こちら, 価格, 高い, 我が家, 大変, 以前, おる
14	0.537	思う, いい, おいしい, 普通, よい, 大きい, いう, 問題 (-1), ほう, 見た目
6	0.527	デザイン, やすい, 手入れ, 使う, コンパクト, サイズ, シンプル, 気に入る, 感じ, 場所
11	0.495	簡単, 飲める, 毎日, コーヒー, 満足, 手軽, くれる, 飲む, 美味しい, 手入れ

単語が見られるが, これも店舗やカタログで写真を見
ることによってデザインやサイズをはじめとした外観の情報
を得ることはできるので, 消費者レビューとして参考
になったと考える閲覧者が他のトピックに比べて少な
くなく考えられる。

表 2 を見ても, 全体的に「参考数」を集めやすい上位
のトピックでは具体的な利用に関する単語が多くみら
れる傾向にある。Filieri (2015) は消費者がレビュー
を評価するとき, 自分のニーズへの関連性, 事実性に
基づいていることを示した。本研究の結果からも閲覧
者が求めている情報が, 実際の利用経験に基づいた具
体的な評価であることが示唆される。

4.3. メーカー別の消費者反応

コーヒーメーカー市場を構成している企業は, それ
ぞれ製品を差別化しているため, メーカーごとにト
ピックが異なる可能性がある。本研究で分析したレ
ビューでは, 全 58 のメーカーについてのレビューが
投稿されている。最も投稿件数が多かったメーカーは
「デロンギ」で 742 件, 続いて「ネスレ」の 655 件,
「パナソニック」の 253 件, 「象印」の 249 件, 「タイ
ガー魔法瓶」の 223 件, 「メリタ」の 212 件と続いて
いる。この上位 6 企業で全レビューの 78.6% を占め
ており, それぞれの企業の戦略と評価を検討する十分
なレビュー数があるため, 本節ではこの 6 つの企業の
所属トピックと「参考数」の集めやすさとの関係を考

察する。

図 4 は, レビューの θ_{dk} を企業ごとにまとめて算
術平均を取り, それを全レビューをまとめて計算した
 θ_{dk} の算術平均で除した値である。ある企業のあるト
ピックについて, 1 よりも大きければ, コーヒーメー
カーカテゴリー全体のトピック出現率と比較して, 当
該企業のレビューで出現する可能性が高いといえる。
また, 1 よりも小さければ, 全体と比較して当該企業
のレビューで出現する可能性が低いことになる。ま
た, 図 4 のトピックは, 参考にされやすさ ρ を基準と
して左から降順に並べている。これは前節の表 2 と対
応している。

図 4 からわかるように, 相対的に参考にされやす
い上位のトピックにおいて, 企業ごとに差異が出てい
る。ただし, 最も「参考数」を集めやすい Topic No.4
については, 全てのメーカーでおおよそ等しく出現し
ている。Topic No.4 は故障や交換に関する内容がま
とめられているため, 全ての企業で共通する話題であ
るといえる。ただし, Topic No.4 以降のトピックで
は企業によって傾向が異なる。Topic No.17 はとく
にパナソニックのレビューに関連する傾向が強く, パ
ナソニックのレビューであれば全体レビューの 3 倍
以上 (314%) の出現率であった。逆に, ネスレのレ
ビューに Topic No.17 は出現しにくく, 全体と比較
して 25.3% にとどまっている。続いて, Topic No.5
は, タイガー魔法瓶のレビューで出現する傾向がやや

図 4. 企業ごとのトピック所属確率

強く、その次に参考にされやすい Topic No. 10 では、ネスレのレビューで出現率が高い。

このように、企業ごとにレビューの傾向が異なり、それぞれの企業の製品について閲覧者に求められている内容が異なることがわかれば、企業のコミュニケーション戦略や、顧客関係マネジメントに活用することができるだろう。

5. 結論と今後の課題

本研究は、インターネット上の消費者レビューを分析対象として、自然言語で記述されるレビュー本文に加えて、0 以上の整数値で記録される閲覧者によるレビューの評価である「参考数」を同時に分類するトピックモデルである「文書評価付き LDA」を提案し、実データへ活用した。本研究のモデルは文書の出現単語 w_i と参考数 y_d が対等であり、どちらもパラメータ θ を事前分布として共有している構造である(図 1)。そのため、文書内で出現した単語のトピック所属確率も、その文書の「参考数」が所属するトピックの影響を受けているという特徴がある。ここが既存の付加情報を用いた拡張トピックモデル(たとえば、Blei & Jordan, 2003; Ramage, Hall, Nallapati, & Manning, 2009)との相違点である。

分析結果からは、「参考数」を集めやすいトピックと集めにくいトピックを分類することができ、両者の内容に質的な差異があることが明らかになった。また、企業別の分析においても、「参考数」を集めるレビューに差異があることが明らかになった。本研究で提案した「文書評価付き LDA」を用いることで、自

然言語を分析対象とした LDA だけでは得られない、閲覧者にとって相対的に重要であると評価されるレビューとそうではないレビューを分類することができるようになった。これを使うことで、レビューサイトの運営者であれば、閲覧者に評価されやすいレビューを投稿することを支援するような具体的な施策を検討することができ、また、レビュー対象の製品を製造するメーカーであれば、現在の利用者がどのような問題に直面しているか、あるいは潜在的な購入者がどのような問題に興味を持っているのかを予測することができ、顧客コミュニケーションに活用することができるだろう。

最後に、本研究で提案したモデルの今後の課題について議論したい。本研究で提案したモデルは、予測の観点からは改善が必要であり、適切なモデルの拡張を行うことで予測能力を向上させることが必要である。モデルの拡張について、LDA についてはこれまでも多くの拡張モデルが提示されており、こうしたモデルを組み込むことで、より活用範囲が広がることが期待される。たとえば Blei & Lafferty (2006) による動的な要因を含めたトピックモデルや、Blei & Lafferty (2007) によって提案されているようなディリクレ分布のパラメータに事前構造を含めたモデルなどがある。また、Li & MaCallum (2006) においては、階層的なトピックを仮定する Pachinko Allocation Model が提案されており、これは Collapsed Gibbs sampling によってパラメータが推定されるため、容易にモデルに組み込むことができるだろう。

次の課題は、評価数と時間の関係についてである。本研究で分析対象となったレビューデータについて

は、投稿日から日が経つほど「参考数」が単調に増加してくことが想定されるため、投稿から日が浅いレビューについては「参考数」が少ない可能性がある。また、一定程度日数が経過すると、製品自体が陳腐化するため、「参考になった数」はある程度の時間が経過すると増えなくなると考えられる。このような性質を反映したモデルの拡張が必要であろう。

謝 辞

本論文の出版に当たり、2名の匿名の査読者より大変有益なコメントをいただきました。また、本研究について、一橋大学の吉岡（小林）徹先生より貴重なコメントをいただきました。この場をお借りして御礼申し上げます。なお、本研究は科研費（17H02573）の助成を受けています。

参 考 文 献

- Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science*, **35**, 6, 953–975.
- Balducci, B., & Marinova, D. (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science*, **46**, 4, 557–590.
- Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2019). Uniting the tribes: Using text for marketing insight. *Journal of Marketing*, **84**, 1, 1–25.
- Blei, D. M., & Jordan, M. I. (2003). Modeling annotated data. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 127–134. Association for Computing Machinery.
- Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, 113–120. Association for Computing Machinery.
- Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, **1**, 1, 17–35.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, **3**, 993–1022.
- Chen, Z., & Lurie, N. H. (2013). Temporal contiguity and negativity bias in the impact of online word of mouth. *Journal of Machine Learning Research*, **50**, 4, 463–476.
- Dabholkar, P. (2006). Factors influencing consumer choice of a 'Rating Web Site': An experimental investigation of an online interactive decision aid. *Journal of Marketing Theory and Practice*, **14**, 4, 259–273.
- Eslami, P., Ghasemaghahi, M., & Hassanein, K. (2018). Which online reviews' do consumers find most helpful?: A multimethod investigation. *Decision Support Systems*, **113**, 32–42.
- Filieri, R. (2015). What makes online reviews helpful? A diagnosticity-adoption framework to explain informational and normative influences in e-WOM. *Journal of Business Research*, **68**, 6, 1261–1270.
- Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, **101**(suppl 1), 5228–5235.
- Humphreys, A., & Wang, R. J. (2018). Automated text analysis for consumer research. *Journal of Consumer Research*, **44**, 6, 1274–1306.
- 岩田 具治 (2015). トピックモデル. 講談社.
- Jiang, Z., & Benbasat, I. (2004). Virtual product experience: Effects of visual and functional control of products on perceived diagnosticity and flow in electronic shopping. *Journal of Management Information Systems*, **21**, 3, 111–147.
- Katsumata, S., & Kim, S. (2020). The text-score allocation model: Finding latent topics of online review documents and multi-item ratings. Discussion Papers in Economics and Business, Graduate School of Economics, Osaka University 20-01, http://www2.econ.osaka-u.ac.jp/econ_society/dp/dp.html.
- Krippendorff, K. (2012). Content analysis: An introduction to its methodology (3rd ed.). Thousand Oaks, CA: Sage Publications.
- Kudo, T., Yamamoto, K., & Matsumoto, Y. (2004). Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, 230–237. Association for Computational Linguistics.
- Lamberton, C., & Stephen, A. T. (2016). A thematic exploration of digital, social media, and mobile marketing: research evolution from 2000 to 2015 and an agenda for future inquiry. *Journal of Marketing*, **80**, 6, 146–172.
- Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, 577–584. Association for Computing Machinery.
- Liu, X., Burns, A. C., & Hou, Y. (2017). An investigation of brand-related user-generated content on Twitter. *Journal of Advertising*, **46**, 2, 236–247.
- López-López, I., & Parra, J. F. (2016). Is a most helpful eWOM review really helpful? The impact of conflicting aggregate valence and consumer's goals on product attitude. *Internet Research*, **26**, 4, 827–84.

- Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., & McCallum, A. (2009). Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 880–889. Association for Computational Linguistics.
- Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly*, **34**, 1, 185–200.
- Pan, Y., & Zhang, J. Q. (2011). Born unequal: A study of the helpfulness of user-generated product reviews. *Journal of Retailing*, **87**, 4, 598–612.
- Poston, R., & Speier, C. (2005). Effective use of knowledge management systems: A process model of content ratings and credibility indicators. *MIS Quarterly*, **29**, 2, 221–244.
- Racherla, P., & Friske, W. (2012). Perceived ‘usefulness’ of online consumer reviews: An exploratory investigation across three services categories. *Electronic Commerce Research and Applications*, **11**, 6, 548–559.
- Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 248–256. Association for Computational Linguistics
- Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*, **51**, 4, 463–479.
- Wedel, M., & Kannan, P.K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing*, **80**, 6, 97–121.
- (2020年4月17日受理, 2020年8月6日採択)
(この間審査3回・審査期間合計74日)