6W-03

BERTによる特徴抽出を駆使した商品レビュー分析モデルに関する一考察

山下 皓太郎* 早稲田大学*
雲居 玄道† 早稲田大学†
蓮本 恭輔‡ 早稲田大学‡
後藤 正幸§ 早稲田大学§

1. 背景と研究目的

近年，EC サイトなど一般消費者が閲覧可能なサイト上で，商品に対するレビューが大量に投稿されている．これらのレビューデータは製品を購入，使用している顧客の感想や要望が記述されており，他の顧客の購買行動に直接的な影響を与える．そのため，メーカーにとって，レビューデータを分析することは顧客のニーズ把握のみならず，既存製品の問題点などを把握するための重要な情報源となっている．しかし，投稿者の書き方には統一性がなく，文体や語彙のばらつきが大きい文書集合の分析となるため，単語別の統計情報のみでは内容の判断がつかないことがある．そこで従来，語彙の類似性の知識表現には，シソーラスなどの辞書が用いられていきた．しかし，シソーラス上の語彙間類似度のみでは表現能力が乏しい．これに対し，近年は大規模な文書データから学習済みのモデルを活用した BERT[1]と呼ばれる自然言語処理技術の有効性が指摘されている．そこで，本研究では BERT を用いて，大量に存在する統一性のない商品のレビューデータから，本当に着目すべき，マーケティング活動に資するレビューの抽出を行う．具体的な事例として，ショッピングモール型 EC サイト A に出品する，大手メーカーB 社の主要商品のレビューデータに提案手法を適用し，その有用性を示す．

2. 問題設定

メーカーにおいて，大量に存在するレビューデータの中から着目すべきレビューを抽出することは，現在の商品に発生している問題の特定や新商品開発の方針を決定する上で重要である．本研究が対象事例とするメーカーB 社では，想定しているラベルと，それらに対応する単語リストを事前に定義し，人手でラベルの付与を行なっている．この手法は，単語のみで判断されるため，リストにない類義語など文書全体の内容は加味されない．また，想定したラベル以外は考慮されないという問題もある．そのため，どの既存ラベルも付与されないレビューが大量に存在している．これらの既存ラベルが付与されないレビューの中には，全体のレビューデータの中に数件しか存在しない特異な話題に関連するレビューも存在する．これらのレビューは，事前にラベルに対応する学習データを用意することも難しく，従来の文書分類の枠組みでは抽出が難しい．しかし，このような特異なレビューは，直近で発生した問題点の早期発見などのために，メーカーにとっては価値のある情報である可能性が高い．

A Study of Analytical Model of Product Review Data Based on Features Extracted by BERT
＊Koutarou Yamashita・WASEDA University
†Gendo Kumoi・WASEDA University
‡Kyosuke Hasumoto・WASEDA University
§Masayuki Goto・WASEDA University

3. 提案手法

3.1 概要

本研究は，商品に対するレビューデータを分析し，レビューデータ全体の傾向把握，および特異なレビューの抽出を行う手法の開発を目的とする．既存の方法では，単語の有無でラベルが付与されているため，想定しているラベルの内容を含むレビューであってもラベル付与されない場合があるという問題があるが，既存のラベルを基にしたマルチラベル分類器の開発によりこれを解決する．また，異常検知の手法を用い，レビューデータ全体に対して他のレビューとは類似性の低い特異なレビューの抽出を可能とする．その上で，この 2 つの手法を組み合わせることで，膨大なレビューデータの中から本来着目すべきで価値が高いと考えられるレビューの抽出を可能とする手法の提案を行う．

3.2 ラベル未付与レビューへのラベル付与

レビューに適切なラベルを付与するマルチラベル問題に対し，ラベルごとの二値分類器を構築し，ラベル付与を行うことを考える．また，本研究では大規模文書データにより事前学習済みの BERT を用いることで，レビュー文を効果的な特徴ベクトルに変換することが可能である．

ここで，ラベル付与されているサイズNのレビューデータ集合$\mathcal{D}^{\mathrm{L}}$とサイズMのラベル未付与データ集合$\mathcal{D}^{\mathrm{T}}$がある．ここで$\mathcal{D}^{\mathrm{L}} \cup \mathcal{D}^{\mathrm{T}}$中の$n$番目のレビューの BERT によって得られた特徴ベクトルを$\boldsymbol{x}_n \epsilon \mathbb{R}^D$とする．このとき，$\mathcal{D}^{\mathrm{T}}$に含まれるラベル未付与レビューにラベル$l$を付与するためのアルゴリズムを以下に示す．

Step1) ラベルが付与されている$\mathcal{D}^{\mathrm{L}}$に含まれる$\boldsymbol{x}_n$を説明変数，ラベルlの有無を表すy_l^nを目的変数として，各ラベル$(l = 1,2,\dots,L)$に対し，それぞれ二値分類器を構築する．

Step2) 構築した分類器を用いて，$\mathcal{D}^{\mathrm{T}}$に含まれるラベル未付与レビューに対し，ラベル予測を行う．

3.3 特異レビュー抽出

特異レビューの抽出においても，BERT によって

得られた特徴ベクトル$\boldsymbol{x}_n$を用いる．特異レビューの抽出は，異常検知手法の 1 つである One Class Support Vector Machine (OCSVM)[2]を用いる．このとき，OCSVM はデータ集合全体に対して特異なデータを検出する教師なし学習手法であることから，全レビューデータ$\mathcal{D}^{\mathrm{L}} \cup \mathcal{D}^{\mathrm{T}}$に含まれる特徴ベクトル$\{\boldsymbol{x}_n\}_{n=1}^{N+M}$を入力することによって実現する．OCSVM は，$n$番目のレビューが特異か否かを表す離散変数$\hat{e}_n \in \{-1,1\}$を出力する．ただし，$\hat{e}_n = 1$は非特異，$\hat{e}_n = -1$は特異なレビューを表し，本稿では特にこの$\hat{e}_n = -1$であるレビューを特異レビューと定義する．

3.4 着目レビュー抽出

3.2 節のラベル付与によって，既存のラベルがいずれも付与されないレビューがある．これは，どのラベルの要素も持たない想定外のレビューであると考えられる．そこで，本研究は価値が高いと考える，メーカーの想定外かつ特異なレビューの抽出を行う．このため，3.2 節の予測値$\hat{y}_l^n$がすべてのlにおいて 0 となるレビューを想定外レビューと定義する．そして，このような想定外レビューのうち，3.3 節により特異レビューにも含まれたレビューを，本研究で最も価値が高いと考える着目レビューと定義し，抽出を行う．

4. 実データ分析

4.1 分析データおよび分析条件

メーカーB 社の主力商品であるレーザープリンタに対するサイト A 上の英文レビューデータに適用する．データ期間は 2017 年 1 月～2021 年 7 月，データ件数は$M + N =$ 19,025 件（うち$\mathcal{D}^{\mathrm{L}}$に含まれるレビュー：$N =$ 13,101 件，$\mathcal{D}^{\mathrm{T}}$に含まれるレビュー：$M =$ 5,924 件）であり，BERT の事前学習モデル[1]より獲得する特徴ベクトルの次元数は$D = 768$，分類器は SVM[3]， 4 種類のラベル Function，Quality， Setup， Toner ごとに分類器を構築する．

評価指標は AUC を用い，$\mathcal{D}^{\mathrm{L}}$のうち 75%を学習，25%をテストデータとして評価した．また OCSVM の事前パラメータ異常値割合は 0.01 とした．SVM，OCSVM ともに，カーネルには RBF カーネルを用いる．

4.2 分析結果

まず，4 種類のラベルにおける分類器の AUC と，分類器を用いてラベル未付与レビューに対するラベル予測を行った結果を表 1 に示し，さらにラベル付与個数ごとのレビュー件数を表 2 に示す．

表 1：各分類器の AUC とラベル付与レビュー件数

ラベル名	Function	Quality	Setup	Toner
AUC	0.818	0.874	0.850	0.910
ラベル付与レビュー件数	941	974	1,109	991

表 2：ラベル付与個数ごとのレビュー件数

ラベル付与個数	0	1	2	3	4
レビュー件数	2,279	3,323	282	32	8

表 1 より，どのラベル分類器においても AUC の値は高く，BERT と二値分類器を用いることで現状の方法ではラベル付与されていないレビューにも正しくラベル付与ができていると考えられる．また表 2 より，ラベル未付与レビューの半数以上のレビューに新たにラベルが付与され，さらにその中でも単一のラベルのみが付与されるレビュー件数が最も多かった．このことから，ほとんどのレビューが単一の内容に関する投稿であると示唆される．

また，ラベル未付与レビューにおいて，分類器によるラベル予測によってもいずれのラベルも付与されなかった想定外レビュー（表 2 においてレビュー付与個数が 0 のレビュー），異常検知手法によって特異と判断された特異レビュー，そしてそれらの双方に含まれる，想定外かつ，特異と判断された着目レビューをそれぞれ分析した．それぞれの分析において抽出されたレビュー件数を表 3 に示す．

表 3：特異レビューとラベル付与との関係性

想定外レビュー件数	特異レビュー件数	着目レビュー件数
2,279	203	119

表 3 より，着目レビューは特異レビューの半数以上であった．このことから着目レビューには他のレビューとは異なる特徴をもったレビューが多く，また想定外レビューであることからも，新たな話題に関連し，詳細に内容を把握するべき話題が含まれていると考えられる．一方で，想定外レビューの多くは特異レビューとはならなかった．これは，想定外レビュー内で類似のレビューが複数含まれている可能性を示唆している．そのため，このようなレビューには，新たなラベルを作成し，付与を行う必要があると考えられる．

5. まとめと今後の課題

本研究では，価値が高いと考えられる想定外かつ特異なレビューの抽出を行い，膨大なレビューデータから本当に着目すべきレビューの抽出が可能な手法の提案を行い有効性を示した．この結果，想定外レビューにも類似のレビュー群が存在する可能性も示された．これは，本研究で用いた 4 種のラベル以外にも B 社ではラベル付与を行っていることからも明らかである．今後の課題として，製品に対する評価値の活用が挙げられる．

参考文献

[1] Devlin, Jacob, et al. ″Bert:Pre-training of deep bidirectional transformers for language understanding.″ *arXiv preprint arXiv:1810.04805*, 2018.

[2] Schölkopf, Bernhard, et al. ″Support vector method for novelty detection.″ *NIPS*. Vol. 12, pp. 582-588, 1999.

[3] Vapnik, Vladimir. ″Pattern recognition using generalized portrait method.″ *Automation and remote control 24*, pp. 774-780, 1963.