D-020

機械学習による商品レビューの文の役割の分類
Classifying Role of Sentence in Product Review by Machine Learning

民岡 佑規† 湯本 高行† 新居 学† 上浦 尚武†
Yuki Tamioka Takayuki Yumoto Manabu Nii Naotake Kamiura

1. はじめに

近年，インターネット上での商取引が普及し，商品のレビューを閲覧する機会が増えている．レビューは主に，商品を購入した消費者が投稿者となり，商品についての情報を記し，投稿したものである．レビューを閲覧することで，商品の具体的な情報が得られ，レビュー閲覧者が商品を購入する際の参考となる．

レビューには，投稿者による商品に対しての感想や出来事などを記した文を含むものがある．他にも，商品の情報を記した文のみで構成されているレビューや，商品の利点と欠点を記した文を含むレビューなど，様々なレビューが存在する．

これらのレビューに記されている文は，役割別に大きく3つに分けることができる．1つ目は投稿者の感想や意思を伝える文，2つ目は商品の情報のみを伝える文，3つ目は投稿者の商品についての出来事を伝える文である．閲覧者は，自身が必要としている情報を伝える役割をもつ文から，知りたい情報を知ることができる．しかし，投稿されたレビューは大量に存在し，その 1 つ 1 つに目を通すことは困難である．閲覧者は，そのようなレビューから必要としている情報を伝える役割をもつ文を探す必要がある．

そこで，本研究ではレビュー閲覧者が必要としている情報を得るために，文がもつ役割を 3 つ定義し，レビュー中における文に対して 3 つの役割のどれに属しているのかの分類を行う．これにより，大量に存在するレビューから，必要としている情報を伝える役割をもつ文のみを閲覧することが可能となる．

現在，商品レビューに対する研究は数多く進められている．関連研究として，岩井ら[1]による，レビュー文に対する文書構造と複数のクラス情報を考慮したあらすじ分類がある．この研究は，小説である「ハリー・ポッター」などのストーリーをもつ商品のレビュー中に，ストーリーの内容を明らかにするあらすじ部分が存在する場合，それらのあらすじを除去することを目的としている．本研究では，あらすじに代わり，家電などのストーリーのない商品を対象とし分類を行う．目的は異なるがアプローチは似ており，構築する分類器に使用する素性は，この研究に使用された素性を参考とする．

本研究の応用として，分類後の文に対する極性の判別が考えられる．関連研究として Hu ら[2]による，商品レビューに対する意見文の抽出と要約の生成がある．この研究では，レビュー中の特徴語から意見語を抽出し，特徴語と意見語を用いて，意見文に対してポジティブかネガティブかの極性を判定する．また，レビューの要約を生成することによって，大量に存在するレビューから必要最低限な情報を表示することを目的としている．本研究で極性の判別を行う場合，たとえばレビュー投稿者の感想が記されている文から，投稿者が商品に対して良いイメージもしくは悪いイメージを抱いているのかを判別することができると考えられる．

†兵庫県立大学, University of Hyogo

2. 役割の定義

レビュー文がもつ 3 つの役割を定義し，クラス分けを行う．各クラスを「意見」，「説明」，「エピソード」クラスとし，人手により各クラスに属するか否かのラベル付けを行う．なお，レビュー中の各文は必ずいずれかのクラスに属するものとする．また，以下に示す条件に複数当てはまれば，重複してクラスに属するものとする．

2.1 意見クラス

商品に対するレビュー投稿者の感想や，レビュー投稿者の意思が記されている文を意見クラスとする．たとえば，「この掃除機は，見た目以上に軽いです．」という文は意見クラスに属する．レビュー閲覧者が商品の購入者の意見や感想を知りたい場合，意見クラスに属する文を閲覧することで必要な情報を探すことができる．

2.2 説明クラス

個人的な感想以外で商品の仕様が記されている文を説明クラスとする．たとえば，「この液晶テレビは32インチです．」という文は説明クラスに属する．また，レビューの中には見やすいように「メリット」と「デメリット」や，「値段」と「サイズ」などのセクションを設け，その下に箇条書きで商品の特徴を述べるものがある．こういった構成は商品の仕様を述べているものが多いため，セクション部分は説明クラスとして扱う．たとえば，「値段・・・75,000 円．サイズ・・・32 インチ．機能・・・番組を検索できる機能付き．」という文は説明クラスに属する．レビュー閲覧者が，商品の具体的な仕様を知りたい場合，説明クラスに属する文を閲覧することで必要な情報を探すことができる．

2.3 エピソードクラス

商品に対する購入者の出来事や投稿者の周りの環境が記されている文をエピソードクラスとする．たとえば，「購入して以来，この音楽プレーヤーを何回も落としてしまっていますが，問題なく動いています．」という文はエピソードクラスに属する．レビュー閲覧者が，自分と似た環境の消費者と商品の間で起こった出来事を知りたい場合，エピソードクラスに属する文を閲覧することで必要な情報を探すことができる．

3. SVM によるクラス分類

本研究では，各クラスの分類に Support Vector Machine（以下，SVM）[3]を使用する．各クラスに分類器を構築

し，計 3 つの分類器を用いる．各分類器に用いる素性は，形態素を基本とし，品詞，レビュー中の文の位置（以下，位置情報），レビューの文の数（以下，文数情報），他クラス分類器の分類結果とする．他クラス分類器の分類結果とは，分類しているクラス以外のクラス分類結果を指す．他クラス分類器の分類結果は，そのクラスに属するか否かという 2 値の素性である．

3.1 意見クラス分類器

意見クラス分類器には表 1 に示す 6 つの素性を用いる．本稿では，この 6 つの素性を意見クラスの基本素性と呼ぶ．

表 1　意見クラスの基本素性

内容	形態素
投稿者の心境を思わせる形態素	満足, 気に入る, 嬉しい
感想を思わせる形態素	思う, 感じ, 感想, 個人的
評価を思わせる形態素	いい, 良い, 悪い, 残念
投稿者の口調を表す形態素	ね
レビュー中の最初の文かどうか	
レビュー中の最後の文かどうか	

表 1 に示した各内容の形態素のいずれかが文に存在している場合，素性が存在するものとする．たとえば，「とても良い商品で満足しています．」という文は「投稿者の心境を思わせる形態素」と「評価を思わせる形態素」をもつ文である．説明，エピソードクラスにおいても同様に素性の有無を決定する．

形態素に加えて，位置情報を用いる．位置情報を用いる理由として，レビューの最初には投稿者の購入に至る経緯，つまり意見クラスに属さない文が多い傾向があると考えられるためである．また，レビューの最後には投稿者の感想，つまり意見クラスに属する文が多い傾向があると考えられるためである．

3.2 説明クラス分類器

説明クラス分類器には表 2 に示す 4 つの素性を用いる．意見クラスと同様に，この 4 つの素性を説明クラスの基本素性と呼ぶ．

表 2　説明クラスの基本素性

内容	形態素
過去を思わせる形態素	た, 今まで, 以前
商品の情報を思わせる形態素	性能, デザイン, 機能, 仕様
説明に用いられづらい形態素	「の」以外の格助詞
レビューに含まれる文の数	

形態素に加えて，品詞，文数情報を用いる．文数情報を用いる理由として，説明クラスに属する文を含むレビューでは，商品の情報を連ねる形で記されていることが多く，結果として文の数が多くなる傾向があると考えられるためである．

3.3 エピソードクラス分類器

エピソードクラス分類器には表 3 に示す 7 つの素性を用いる．意見クラスと同様に，この 7 つの素性をエピソードクラスの基本素性と呼ぶ．

表 3　エピソードクラスの基本素性

内容	形態素
過去を思わせる形態素	た, 今まで, 以前
投稿者の情報を思わせる形態素	私, 僕, 我が家
感情を表す形態素	!, ！
周りの環境を思わせる形態素	私の, 僕の, 用
具体的な内容を思わせる品詞	数詞
レビュー中の最初の文かどうか	
レビュー中の最後の文かどうか	

形態素に加えて，品詞，位置情報を用いる．位置情報を用いる理由として，意見クラスと同様に，レビューの最初には投稿者の購入に至る経緯，つまりエピソードクラスに属する文が多い傾向があると考えられるためである．また，レビューの最後には投稿者の感想，つまりエピソードクラスに属さない文が多い傾向があると考えられるためである．

3.4 他の分類器の結果の利用

本研究では，最も分類精度が良かったクラス分類器を選択し，その分類結果を素性として残りのクラス分類器に使用する．また，2 番目に分類精度が良かったクラス分類器も分類器として適切であれば，その分類結果を素性として残りのクラス分類器に使用する．

たとえば，実験を行った結果，エピソードクラス分類器の分類精度が最も良く，説明クラス分類器の分類結果が最も悪かった場合，まず，素性としてエピソードクラス分類器の分類結果を意見，説明クラス分類器に追加する．次に，2 番目に分類精度が良い意見クラス分類器が分類器として適切である場合，素性として意見クラス分類器の分類結果を説明クラス分類器に追加する．なお，本研究では分類器としての適切さは，ランダム分類に勝るか否かによって決定する．

4. 実験

本研究では，形態素解析には MeCab[4]を使用し，SVM による分類には LIBSVM[5]を使用する．カーネルは RBF カーネルを使用し，分類の際に最適なパラメータはグリッドサーチにより決定する．また，各クラスで 5 分割交差検定を行い，適合率，再現率から F 値を算出することで分類精度の評価を行う．各値を求める式を以下に示す．

$$\text{適合率} = \frac{\text{人手が正例かつ分類器が正例とした数}}{\text{分類器が正例とした数}}$$

$$\text{再現率} = \frac{\text{人手が正例かつ分類器が正例とした数}}{\text{人手が正例とした数}}$$

$$\text{F 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}}$$

4.1 データセット

Amazon.co.jp における 10 商品についてのレビューから，ランダムに 10 件ずつ取得する．実験に用いる 10 商品の内容を表 4 に示す．

表 4　実験に用いる商品

DVD プレーヤー	カーテン
カメラ	ソファ
テレビ	ヘッドホン
掛け時計	空気清浄機
掃除機	本棚

取得したレビューの文の数の合計は 675 である．各クラスの正例（クラスに属する文）と負例（クラスに属さない文）の数を調べ，表 5 に示す．

表 5　データセット

クラス	正例の数	負例の数
意見	398	277
説明	244	431
エピソード	254	421

各クラスにおいて正例と負例の数に偏りがみられる．本研究では，使用するデータに対して，サンプリングを行うことでデータの均一化を図る．サンプリングは正例と負例を比較し，数が小さいほうに合わせる．

4.2 クラス分類器の順番の検討

最初に，素性として他クラス分類器の分類結果を使用するために，基本素性を使用した場合における各クラス分類器の分類精度の比較を行う．得られた結果を表 6 に示す．

表 6　各クラス分類器の分類精度

クラス分類器	適合率	再現率	F 値
意見	0.781	0.477	0.592
説明	0.790	0.865	0.826
エピソード	0.683	0.331	0.446

分類精度を比較したところ，説明クラス分類器の F 値が最も大きいことがわかる．よって，素性として説明クラス分類器の分類結果を意見，エピソードクラス分類器に使用することにより分類精度の向上を図る．

意見，エピソードクラス分類器の素性に，説明クラス分類器の分類結果を追加し，評価を行う．また，説明クラス分類器の分類精度が 100%，つまり人手による分類である場合についても評価を行う．

まず，意見クラス分類器について評価を行う．得られた結果を表 7 に示す．また，表 7 の関係を表すグラフを図 1 に示す．

表 7　意見クラス分類器の分類精度
（説明クラス分類器の分類結果を使用）

素性	適合率	再現率	F 値
基本素性	0.781	0.477	0.592
基本素性＋ 説明分類器の分類結果	0.766	0.484	0.593
基本素性＋ 人手による説明の分類	0.781	0.477	0.592

意見クラス分類器については，素性追加前と素性追加後で大きな F 値の変動がみられない．このことから，説明クラス分類器の分類結果は F 値の改善に貢献していないと考えられる．

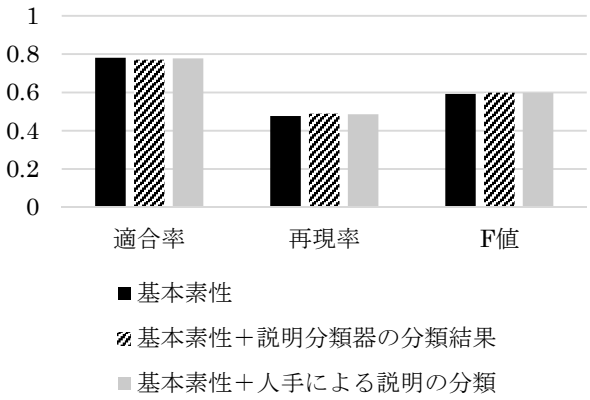

図 1　意見クラス分類器の分類精度
（説明クラス分類器の分類結果を使用）

次にエピソードクラス分類器について評価を行う．得られた結果を表 8 に示す．また，表 8 の関係を表すグラフを図 2 に示す．

表 8　エピソードクラス分類器の分類精度
（説明クラス分類器の分類結果を使用）

素性	適合率	再現率	F 値
基本素性	0.683	0.331	0.446
基本素性＋ 説明分類器の分類結果	0.669	0.811	0.733
基本素性＋ 人手による説明の分類	0.634	0.846	0.725

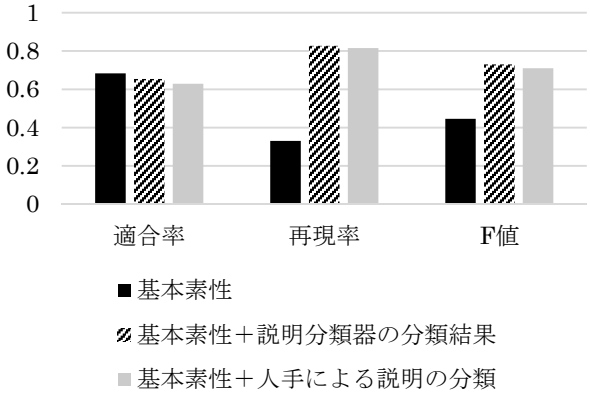

図 2　エピソードクラス分類器の分類精度
（説明クラス分類器の分類結果を使用）

エピソードクラス分類器については，素性追加前と素性追加後で適合率に大きな変動がみられないが，素性追加後のほうが追加前よりも高い F 値となっている．このことから，説明クラス分類器の分類結果は素性として有用であると考えられる．また，説明クラス分類器と人手の分類結果を比較すると，F 値の低下がみられる．この結果から分類結果が正しいことが必ずしも分類精度の向上にはつながらないと考えられる．

次に，素性としてエピソードクラス分類器の分類結果を意見クラス分類器に使用することで分類精度の向上を図る．

なお，エピソードクラス分類器に使用する素性は基本素性に加えて，説明クラス分類器の分類結果とする．得られた結果を表 9 に示す．また，表 9 の関係を表すグラフを図 3 に示す．

表 9　意見クラス分類器の分類精度
（説明，エピソードクラス分類器の分類結果を使用）

素性	適合率	再現率	F 値
基本素性	0.781	0.477	0.592
基本素性＋ 説明分類器の分類結果	0.766	0.484	0.593
基本素性＋ 説明，エピソード分類器 の分類結果	0.766	0.484	0.593

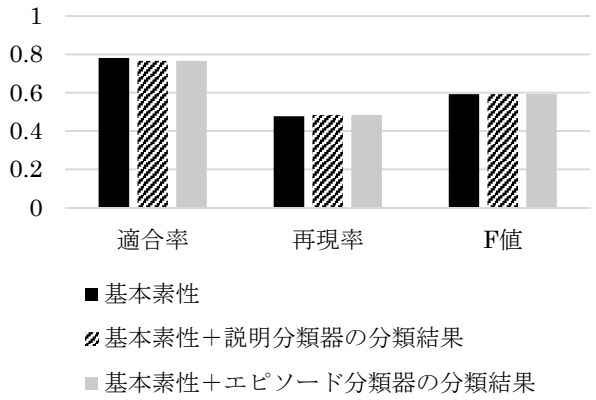

図 3　意見クラス分類器の分類精度
（説明，エピソードクラス分類器の分類結果を使用）

素性としてエピソード分類器の分類結果を追加する前と後で F 値に変動がみられない．このことから意見クラス分類器において，説明，エピソードクラス分類器の分類結果という素性は分類精度の向上に貢献していないと考えられる．

各クラス分類の結果として，意見，説明クラス分類においては基本素性，エピソードクラス分類においては基本素性に加え，説明クラス分類器の分類結果を用いることで最も分類精度が良くなることがわかった．分類する際の各クラスの順番を図 4，最終的な分類精度を表 10 に示す．

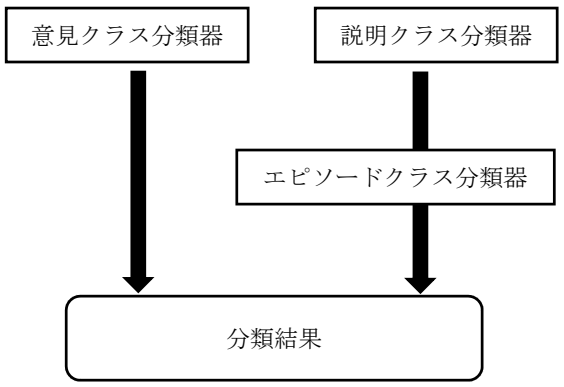

図 4　分類の手順

表 10　各クラス分類器の最終的な分類精度

クラス分類器	適合率	再現率	F 値
意見	0.781	0.477	0.592
説明	0.790	0.865	0.826
エピソード	0.669	0.811	0.733

5. おわりに

本研究では，商品レビューの文に 3 つの役割を定義し，3 つのクラスとして意見，説明，エピソードクラスを定義した．意見クラスは投稿者の意見や意思を伝える文，説明クラスは商品の情報のみを伝える文，エピソードクラスは投稿者の商品についての出来事や投稿者の周りの環境を伝える文である．

また，それらのクラスを分類する分類器として SVM を使用し，各クラスに対して 2 値分類を行う分類器を構築した．素性は，形態素，品詞，位置情報，文数情報，他クラス分類器の分類結果を使用した．他クラス分類器の分類結果については，分類精度が優れているものから順番に選択し，他のクラス分類器に使用した．実験の結果，最も分類精度が良かった分類器は説明クラス分類器，2 番目に良かった分類器はエピソードクラス分類器であった．よって，説明クラス分類器の分類結果を意見，エピソードクラス分類器に使用し，エピソードクラス分類器の分類結果を意見クラス分類器に使用した．結果として，意見クラス分類器については分類精度の向上がみられなかった．また，エピソードクラス分類器については素性の追加によって分類精度が向上し，説明クラス分類器の分類結果は素性として適切であると考えられる．

今後の予定として，係り受け解析を行い，文節間の関係を素性として用いることで分類精度の向上を図る．また，素性として使用している形態素の種類を増やし，辞書を作成することを考えている．

謝辞

本研究の一部は，平成 27 年度科研費若手研究（B）「情報の詳細関係に基づく Web ページの組織化」（課題番号：24700097）によるものである．

参考文献

[1] 岩井 秀成, 土方 嘉徳, 西田 正吾, “レビュー文に対する文書構造と複数のクラス情報を考慮したあらすじ分類”, WebDB Forum (2013).
[2] Minqing Hu, Bing Liu, “Mining and Summarizing Customer Reviews”, KDD (2004).
[3] V.N. Vapnik, “Statistical Learning Theory”, Wiley, New York (1998)
[4] MeCab, ttp://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[5] LIBSVM, http://www.csie.ntu.edu.tw/~cjlin/libsvm/