6Q-09

Twitterにおける口コミ情報の抽出と分析

王　博†　　服部 隆志‡　　萩野 達也‡

慶應義塾大学大学院 政策・メディア研究科†　　慶應義塾大学 環境情報学部‡

1. はじめに

毎日Twitter上で人々が感想や気持ちを他人に伝えるために投稿し，膨大なデータが生まれている．その中から，何かを評価する文書を抽出することができれば，自動的に評判を集約することができるなど，利用する価値がある．先行研究では，指定した特定の対象に対する評価の抽出を行っていたが[1]，本稿では，先行研究を踏まえ，より一般的な対象に対する評価の抽出と分析方法について提案する．本研究では，自然言語処理と機械学習を利用してTwitterの投稿から「評価対象・属性・属性値」の3つ組の評価を抽出・分析し，評価対象の特徴を抽出することを行った．また，評判情報を可視化するために，評価対象の特徴によるレーダーチャートを作成した．

2. データセット作成

Twitterの投稿には様々なトピックがあり，評価情報を含むコンテンツの割合が非常に小さい．本研究では，指定したキーワード「ラーメン」を含む二週間の投稿549,623件から，ランダムで3,060件Tweetを選び，人手でラベル付けを行った．

判定基準は，筆者によって評判情報があると判断した場合，評価表現を持つツイートとしてラベルを付ける．3,060件の中で評価表現があると判断したツイートは625件であった．

	訓練データ	テストデータ	総計
評価表現	491	134	625
一般表現	1970	465	2435
総計	2461	599	3060

表 1　データセット

3. 機械学習による評価ツイートの抽出

評価表現を持つツイートの抽出には，機械学習のアンサンブル学習を用いて判断した．

3.1 テキスト処理

ツイート本文を形態素解析し，名詞，形容詞，動詞を抽出する．検索キーワード「ラーメン」と符号などのストップワードを除いた．出現頻度を統計処理し，上位3,000個の単語を辞書として作成した．ベクトル化したツイートを分類器に入力した．

3.2 分類器の比較

本研究では，線形カーネルSVM分類器（LinearSVC），多項モデルナイーブベイズ分類器（MultinomialNB）とベルヌーイモデルナイーブベイズ分類器（BernoulliNB）を使って評価した．それぞれの分類器の精度を表2に示す．

	分類器		
	LSVC	MNB	BNB
適合率	0.664	0.653	0.759
再現率	0.724	0.701	0.634
F値	0.693	0.676	0.691

表 2　分類器の学習精度

3.3 アンサンブル学習の精度検証

表2の各分類器を弱学習器としてアンサンブル学習を行うために，3つの分類器の何個以上が正例と判定した時に正例と判定するかを比較した．

	正例と判定した分類器の数		
	3個	2個以上	1個以上
適合率	0.867	0.734	0.588
再現率	0.485	0.701	0.873
F値	0.622	0.717	0.703

表 3 アンサンブル学習の精度

表3から，多数決で正例と判定した，分類器の数が2個以上の場合の精度が一番高いことが分る．このため，本研究では，2つ以上の弱学習器が正例と判定した時に正例と判定することにした．全部のツイート549,623件をアンサンブル学習で分類した結果，評価表現を含むツイート59,038件（10.74%）を抽出した．

4. 評価対象の抽出

4.1 店名辞書作成

評価対象を「ラーメン屋」に限定するためには，評価対象辞書を作る必要がある．そのため，飲食店レビューサイト「食べログ」からラーメ

Extraction and Analysis of Online Reviews on Twitter
†WANG BO, Keio University, Graduate School of Media and Governance
‡Takashi HATTORI, Tatsuya Hagino, Keio University, Environment and Information Studies

ン屋の店名 52,000 件を収集した．店名の中から，一つの仮名や漢字のものを削除し，Mecab のユーザ辞書として店名辞書を作成した．

店名辞書において，各店名の重みを変える必要がある．例えば，「濃厚豚骨　まつり」という店名のようにスペースを含む場合，曖昧性がほぼないため一番大きいな重みを付ける．一方，「ありがとう」などラーメン屋の店名以外も意味がある単語に対しては比較的に小さな重みを付けた．

4.2 評価対象抽出

店名の抽出には，ルールベースを用いた．抽出した店名の前後 2 形態素を合わせて抽出し，筆者が設定したルールを使って判定する．表 4 は，評価ツイート 300 件の中で店名が出現する場所に応じて作ったルールである．

形態素 1	'#', 'in', '@','at', '@', '麺屋', 'ラーメン', 'らーめん','店'
形態素 2	'「', '【', '『', '〖', 'at', '@','　', '　','¥n'
店名候補形態素 3	店名辞書にある単語
形態素 4	'#', '麺', 'さん', 'in', 'at', '@','　', '　','¥n'
形態素 5	'#', '麺', 'さん', 'in', 'at', 'ラーメン','らーめん','@'

表 4 対象検出のルール

5. 属性・属性値対の抽出

ルールベースにより判定した店名に対する評価ツイートについて，属性と属性値を抽出する．

本研究では，評価ツイートの係り受け解析を通じて「名詞，名詞・形容詞・動詞」の対を抽出した．高い頻度の上位 1,000 個をラーメン屋に関連する名詞として選択した．本研究では総計 267 個の単語を属性辞書として作成した．そして，属性辞書による「名詞，名詞・形容詞・動詞」の対から属性・属性値を抽出し，一つ評価対象の属性と属性に関連する属性値を集約した．

6. 評判分析・可視化

前述で抽出した店名を集約し，属性・属性値の対における点数付けを行った．「名詞，名詞・形容詞・動詞」の対による，属性の点数は属性値を感情辞書によって判断して付けた．本研究の感情辞書は，乾ら[2]が公開したネガ・ポジ辞書を使って作成した．感情辞書を使って 1 点から 5 点までの範囲で，点数を算出した．次に，一つの評価対象におけるすべての属性・属性値の対を集約し，属性ごとに加重平均点を付けた．また，一つの評価対象が多く属性を持つ場合，属性の点数の大きい順から上位 7 個を選び，対象の特徴とした．

例えば，図 1 は 2 つ店舗がそれぞれ別の属性を特徴に持つことを示している．検証のために，「食ベログ」における各の口コミページで「いいね順」の上位 20 件の口コミを収集し，特徴の属性が出現した回数を集計したものを表 5 に示す．本研究で抽出した評価対象の特徴の属性は「食ベログ」の口コミ文書の中にも頻繁に言及されている．特に「らぁめん鴇」の口コミ文書には，「二郎」の特徴の属性の「ニンニク」と「野菜」と「卵黄」がない．これより，本研究によって，店の特色あるいは注目されているポイントを抽出することができたことが分かる．

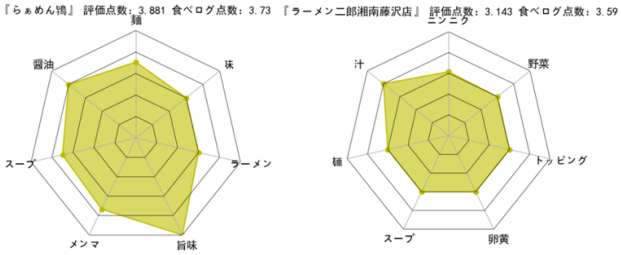

図 1　レーダーチャート

	らぁめん鴇		二郎	
特徴属性	麺	106	ニンニク	106
	味	204	野菜	102
	ラーメン	93	トッピング	32
	旨味	9	卵黄	5
	メンマ	17	スープ	78
	スープ	69	麺	227
	醤油	110	汁	92
参考属性	ニンニク	0	メンマ	0
	野菜	0	醤油	24
	卵黄	0		

表 5「食ベログ」投稿の特徴属性の出現回数

7. おわりに

本研究では，収集した 14 日間のツイートに基づいて，評価表現の抽出，評価対象，特徴属性，属性値の検出を行い，最後に評価情報を可視化した．結果として，本研究では，Twitter における評判情報を抽出・分析する手法を実現した．また，評価対象の特徴を抽出し，可視化することができた．

参考文献

[1] 鳥海 不二夫，Twitter 上のビッグデータ収集と分析，組織科学　48(4)，47-59，2015

[2] 乾 健太郎，日本語評価極性辞書，http://www.cl.ecei.tohoku.ac.jp/index.php?Open%20Resources%2FJapanese%20Sentiment%20Polarity%20Dictionary，東北大学 乾・岡崎研究室，2008