D-012

ソーシャルブックマークを学習データとして用いたツイートの印象の推定
Estimating Impressions of Tweets by Social Bookmarks as Training Data

岡村 康行† 湯本 高行† 新居 学† 佐藤 邦弘†
Yasuyuki Okamura Takayuki Yumoto Manabu Nii Kunihiro Sato

1. はじめに

Web ページを検索し閲覧する際，ページのトピックだけではなく，他のユーザがどのように評価しているかといった評判も重要な判断基準である．

他のユーザの評価が投稿されているサービスとして，ソーシャルネットワーキングサービス（SNS）が挙げられる．このなかでも，Twitter[1]はユーザが 140 文字以内の短文を投稿し，共有できるサービスであり，コンピュータだけでなく携帯端末などからも気軽に投稿できるため，近年では日本国内でも利用者が増加しており[2]，2013 年 11 月時点で利用者は2070万人だといわれている．従って，Twitterでは多くのインターネットの利用者の興味や関心を示すツイートが投稿されている可能性が高いと考えられる．そこで，本研究では，Web ページに関するユーザの意見が記述されているツイートに着目する．しかし，ツイートのみで学習と分類を行う場合，多くは Web ページの評価ではないことや，学習データの選択が容易ではないことが課題となる．

一方，ソーシャルブックマーク（以下 SBM）は Web 上にブックマークを保存するサービスであり，ユーザの興味や関心を集める Web ページが数多く登録されている．コメントだけではなく，内容や感想により付与されたタグを用いることで，比較的容易にページを分類することが可能である．本研究では，日本国内で分野にかかわらず汎用的に使用されており，2012 年 3 月時点で 375 万人の会員[3]が登録している日本最大規模の SBM サービスである，はてなブックマーク[4]のデータを使用する．

本研究では SBM のデータを学習データとして Support Vector Machine で印象の分類器を構築し，これを用いて Web ページの URL が含まれるツイートの分類を行う．具体的には，SBM のタグから教師信号を生成し，コメントから素性ベクトルを生成して学習データとする．対象とする投稿はポジティブ，ネガティブだけではなく，ポジティブと分類されたコメントは，さらに有用な情報か面白い情報かに分類し計 3 種類のクラスに印象を分類する．また，本手法は学習データの選択も機械的に行うため，分類においては人手による処理は不要であることも特徴である．

2. 関連研究

SBM は Web 上にブックマークを保存するサービスであり，ブックマークを分類する方法として，タグと呼ばれるテキストを自由に付与できる方式をとり，更にコメントを入力できることが特徴である．

1 件のブックマークは以下のように表される．

$$\boldsymbol{b} = (u, r, t, c)$$

u はユーザ名(User)，r は URL(Resource)，t はタグ(Tag)，c はコメント(Comment)を表す．

† 兵庫県立大学, University of Hyogo

SBM に関する研究はいくつか行われているが，タグによる分類については Golder ら[5]の研究がある．この研究ではユーザが付与するタグの種類として以下の 7 つの役割を挙げている．

1. 何について書かれたものか（例：料理，PC）
2. それが何なのか（例：記事，本，ブログ）
3. 誰の所有物か
4. これまでに付与したタグを洗練した新カテゴリー
5. 印象や性質（例：これはすごい，便利）
6. 自身に関連するもの
7. 行動に関連するもの（例：あとで読む）

特に 1. に示す役割に着目し，利用者に対する情報の推薦を行う研究には Sen ら[6]や丹羽ら[7]の研究がある．また，SBM を Web 検索に役立てる手法として山家ら[8]の研究がある．この研究では，あるページをブックマークしたユーザ数に着目し，そのページの人気度として使用するが，ユーザがどのような印象を持っているかまでは判断できない．

本研究では，特に 5. に示すユーザの印象や性質などが述べられているタグに着目し分類を行う．

機械学習を用いてツイートの極性を判定する研究では，Go ら[9]がポジティブかネガティブかという分類を行っているが，本研究ではさらにポジティブな情報の中でも Web 検索をする際に必要とされることが多い有用な情報か，あまり必要とされない面白い情報かという分類も行う．

SBM のタグを元に学習し，Twitter を対象に評価を行う研究には，齋藤ら[10]の研究がある．学習や分類に用いるデータは類似しているが，目的が興味語の抽出やユーザの推薦である点が，SBM のコメントやツイートの印象の推定を行う本研究とは異なる．

また，本研究では機械学習アルゴリズムの一種である Support Vector Machine（以下 SVM）を用いる．SVM の学習及び分類は LIBSVM[11]を用い，カーネル関数として RBF を用いる． LIBSVM ではオプションを設定することにより，クラス分類結果だけではなく 0 から 1 までの推定確率[12]を出力できる．そのクラスに属する場合，推定確率として 0.5 以上の値が出力される．

3. 提案手法

本研究では，まず SBM のタグを用いて Web ページに付与されたブックマークコメントのラベル付けを行い，印象の分類器を構築する．次にこの分類器を用いて，ツイートの分類を行う．

3.1 SVM による印象の分類

印象は大きく分けて，ポジティブ，ネガティブに分類することができる．さらに，ポジティブの中でも，有用な情報や面白い情報といった分類が可能である．

たとえば本研究の手法を用いてユーザの印象を集計し情報検索へ応用する場合，一般的に有用な情報が求められて

いる場合が多い．従って，有用な情報のみを提示することによりユーザが求めている情報にたどり着きやすくなることが考えられる．逆にユーザが悪い点を知りたい場合はネガティブな情報を提示し，娯楽目的で検索する際は面白い情報を提示することがよいと考えられる．

従って本研究では以下のように印象のクラスを定義する．

- Positive：ポジティブな情報
  - Useful：有用な情報
  - Funny：ジョークや面白い情報などの娯楽情報
- Negative：ネガティブな情報

2 値分類器である SVM を用いて，4 クラスの分類を行うため，まず Positive と Negative の 2 種類の分類器を用いて分類し，さらに Positive に分類されたものについては Useful と Funny の 2 種類の分類器を用いて分類を行う．この際，推定確率が 2 種類とも 0.5 未満の場合は分類不能として除外する．また，テストデータと学習データにおいて一致する素性が存在しない場合も分類不能として除外する．

上記のアルゴリズムを Algorithm 1 に示す．M_{pos}, M_{neg}, M_{use}, M_{fun} はそれぞれのクラスにおいて学習した SVM モデルであり，その構築方法は 3.3 節に示す．predict 関数は，引数として SVM モデルと一件の入力データ D を与えることで，そのモデルにより分類を行った推定確率 P を出力する関数である．P_{pos}, P_{neg}, P_{use}, P_{fun} はそれぞれのクラスの推定確率である．推定確率の比較を行い，分類結果のクラス C を出力し，いずれにも該当しない場合は other とする．

Algorithm 1 推定確率を用いたクラスの分類

Input : M_{pos}, M_{neg}, M_{use}, M_{fun}, D
Output : C
1: $P_{pos} \leftarrow$ predict (M_{pos}, D)
2: $P_{neg} \leftarrow$ predict (M_{neg}, D)
3: **if** $P_{pos} > 0.5$ AND $P_{neg} < 0.5$ **then**
4: $P_{use} \leftarrow$ predict (M_{use}, D)
5: $P_{fun} \leftarrow$ predict (M_{fun}, D)
6: **If** $P_{use} > P_{fun}$ AND $R_{use} > 0.5$ **then**
7: $C \leftarrow$ useful
8: **else if** $P_{fun} > 0.5$
9: $C \leftarrow$ funny
10: **else**
11: $C \leftarrow$ other
12: **end if**
13: **else if** $P_{neg} > 0.5$ AND $P_{pos} < 0.5$ **then**
14: $C \leftarrow$ negative
15: **else**
16: $C \leftarrow$ other
17: **end if**
18: **return** C

3.2 コメントを用いた特徴ベクトルの作成

ブックマークのコメントおよびツイート（以下コメントと総称）から特定の品詞を抽出するために，形態素解析を行う．本研究では，形態素解析器 Juman[13]により形態素解析を行う．Juman では分割した形態素に対し，読み，品詞だけでなく，代表表記も取得できる．代表表記により表記揺れの問題を取り除くことができ，たとえば「おいしいイチゴ」と「美味しい苺」は同一の表記として扱うことができる．

本手法では，コメント中に出現する印象を表す語に着目する．この印象を表す語を特徴語とする．特徴語は，品詞が名詞，形容詞と判定された語の代表表記と，顔文字として判定された記号を用いる．なお，名詞と分類された語でも数詞は除外した．

コメントにはインターネット特有の表現や顔文字が出現し，形態素解析において十分な精度が得られない場合がある．Juman では標準で Wikipedia のタイトルを用いた辞書や，顔文字辞書を用いた分類が行える．また，コメントではネットスラングで笑うという意味を示す「www」がよく用いられており，特に面白い情報を示す Funny で使われることが多い．たとえば，「良いねぇ wwww」というコメントの場合，形態素解析を行うと wwww は未定義語となり特徴語には該当しないため，以下の条件を満たす場合は特徴語としてネットスラング「www」も加える．なお，w の連続数にかかわらず出現数は 1 回とする．

- コメントの末尾が w
- w が 2 文字以上連続するが，URL の形式ではない

さらに，同一の URL に対して全く同じ内容のコメントが複数件存在する場合，ユーザのコメントではなくページ作成者の指定するコメントの初期値である可能性が高いため除去する．

全てのコメントをコメントリスト$\boldsymbol{v}$で管理し，全てのコメントで共通する特徴語のリストを作成する．コメントを特徴ベクトルに変換すると次式のようになる．

$$\boldsymbol{v_k} = (w_{k1}, \cdots, w_{kl}, \cdots, w_{kN})$$

k はコメントリストの k 番目の項目を表し，l は特徴語のリストの l 番目の特徴語を表す．従って，w_{kl} は k 番目のコメントで l 番目の特徴語が使用されている場合は 1，使用されていない場合は 0 を表し，同一コメントで複数回同じ特徴語が使用される場合でも 1 回の出現と見なす．

ブックマークコメントのベクトル化の過程において，クラスにかかわらず頻繁に使用される語は特徴語として不適切であると考え，多くのコメントで使用されている名詞は除去する．具体的な基準は 4.1.1 項で述べる．

3.3 タグを用いた Web ページの印象推定

人手による学習を必要とせず機械的に Web ページの印象を分類する方法として SBM のタグを用いる．分類された Web ページに付与されたコメントを学習コメントとして選択する．各クラスの分類器を構築する際に，表 1 に示すタグが多く付与されている Web ページから順に正例，負例の学習コメントとして選択する．また，幅広く形態素を学習するため，同一の URL において同一条件に一致するコメントが複数件存在する場合でも，学習コメントとして選択するコメントは 1 件の URL につき 1 件のみする．

この手法により学習コメントのデータセットを全てのクラスに対して作成し，SVM により学習を行い 4 種類の分類器 M_{pos}, M_{neg}, M_{use}, M_{fun} を構築する．

表 1　クラス名と正例負例に対応するタグ一覧

クラス名	正例とするタグ	負例とするタグ
Positive	これはすごい，お役立ち，ネタ	これはひどい
Negative	これはひどい	これはすごい，お役立ち
Useful	お役立ち	ネタ，これはひどい
Funny	ネタ	お役立ち，これはひどい

3.4 ツイートの分類

Twitter の API を用いることで，ツイートの本文，ツイートの言語，URL，ツイートの投稿に用いたクライアントソフトウェア（クライアント）名などが取得できる．

取得したツイートから，まずは日本語の URL 付きのツイートのみを選択する．また，診断系サイトや特定の URL 付きツイートをすることが条件の懸賞目的の投稿は除去する必要がある．さらに，機械的に投稿されたツイートはクライアントソフトウェアより判断できるため，Twitter で頻繁に使用されているクライアントのうち，PC や携帯端末などからユーザが入力し投稿するクライアントに限定した．

ユーザによって投稿された URL 付きツイートのみを分類するため，表 2 に示す条件を全て満たすツイートのみを抽出とする．

表 2　ツイートの選択条件

種類	リプライ，リツイートではない
本文	「ギフト」「招待」「QUO カード」を含まない
URL	診断メーカー(http://shindanmaker.com)以外の URL にリンクする
クライアント	以下のいずれかに該当する ● Twitter 公式クライアント ● Twitter 公式サイト ● Web ページのツイートボタン ● Twicca ● Janetter ● TweetDeck ● ついっぷる ● Ustream.TV

4. 実験

4.1 節では提案手法による学習アルゴリズムに対して，人手によりラベル付けした SBM のコメントに対する正解データを利用して評価を行った，次に 4.2 節では，この手法で構築した分類器を用いて分類を行い，人手によりラベル付けしたツイートを正解データとして評価を行った．

分類結果の評価指標として，再現率，適合率，F 値を用いる．このうち F 値が高いほど分類性能が良いといえる．各指標の定義は以下の通りである．

再現率（Recall） 全ての正解データのうち，分類器が正解として出力した割合

$$Recall = \frac{R_{Q\to Q}}{R_{Q\to *}}$$

適合率（Precision） 分類器がクラスに分類した結果のうち，実際に正解のデータである割合

$$Precision = \frac{R_{Q\to Q}}{R_{*\to Q}}$$

F 値 再現率および適合率の調和平均

$$F = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$$

ここで*は特定のクラス Q に関わらず全てを対象にするワイルドカードであり，$R_{Q\to *}$は正解が Q である全ての分類結果，$R_{*\to Q}$は分類器により Q として分類された全ての数を表す．ここで，正解は Q であるが，分類不能と評価されたコメントについては，計算に含まれない．

提案手法において，まず Positive, Negative に分類し，次に Positive に分類されたものは Useful, Funny の分類を行う．分類器の出力として Negative, Useful, Funny の 3 クラスであるため，Positive の分類精度の計算においては Useful と Funny の和を用いる．

4.1 SBM による印象分類器の構築

SBM のタグを用いて Web ページに付与されたコメントを学習コメントとして分類器を構築し，コメントに対する正解データを用いて精度を評価する．

4.1.1 SBM データの取得

学習データ及びテストデータの元となるブックマークデータとして，SBM サービス上で公開されているブックマーク情報を取得した．本研究では，はてなブックマークをクローリングして作成されたデータを使用する．本実験で使用するデータベースの規模を表 3 に示す．クラスにかかわらず多くのコメントで使用されている語は特徴語として不適切であると考え，2000 件以上のコメントで使用されている名詞は除去した．

表 3　収集した SBM データの規模

ユニーク URL 数	18,687
コメント存在 URL 数	15,876
総コメント数	340,595
取得名詞･形容詞種類数	37,343
ネットスラング（www）を含むコメント数	14,769

また，SBM による分類器の精度を確認するため，SBM のコメントを無作為に選択し，人手によりラベル付けを行った．クラス別のラベル付け数を表 4 に示す．

表 4　人手による SBM コメントラベル付け結果

クラス	コメント数
Useful	1,409
Funny	545
Negative	741

4.1.2 SBM のコメントの分類結果

表 4 のデータを正解データとし，分類を行った．3.2 節で述べた顔文字と笑うを意味するネットスラングの有無で条件を設定し比較を行った．SVM の RBF カーネルを使用する際に設定するパラメータの探索のため，グリッドサーチ

を行い，$C = 2, \gamma = 0.03125$とした．各条件における 4 クラスの F 値を表 5 に示す．表の「顔文字」列は顔文字の有無を示し，「www」列は笑うを意味するネットスラングの有無を示している．各クラスの分類結果において最も F 値が高い結果を太字で示している．

表 5　SBM のコメントの分類結果

顔文字	www	Negative	Positive	Funny	Useful
なし	なし	**71.45**	**85.96**	44.72	73.75
なし	あり	67.13	81.47	54.63	74.71
あり	なし	67.16	83.33	50.88	74.46
あり	あり	66.55	81.32	**54.92**	**74.85**

表 5 より，Funny, Useful の分類においては，顔文字と笑うを意味するネットスラングを含める場合の精度が高いが，Positive, Negative の分類においては除去した場合の精度が高いことがわかる．

また，分類精度が最も高い組み合わせである，Positive, Negative の分類においては顔文字とネットスラングを除外し，Funny, Useful の分類においては顔文字とネットスラングを含めるという条件で分類を行った．このときの再現率，適合率，F 値を表 6 に示す．

表 6　SBM のコメントの分類精度最大時の結果

クラス	Negative	Positive	Funny	Useful
再現率	72.34	85.45	64.34	68.81
適合率	70.59	86.48	47.92	82.15
F 値	71.45	85.96	54.93	74.89

Funny, Useful は Negative, Positive の分類結果に依存するため，分類精度が高い条件を組み合わせることにより，精度が向上したことがわかる．

しかし，SBM のコメントの分類においては Funny の分類精度が他のクラスと比較し低い．表 4 に示すとおり，はてなブックマークにおいては Funny と比較し，Useful のコメントが多いだけでなく，Funny の分類に失敗したコメントでは特徴語だけ閲覧すると分類が困難であるコメントが多いことも原因と考えられる．例を以下に示す．[]内に示す語が抽出された特徴語である．

- 無理！絶対無理ですぅぅ＞＜！ [無理]
- カメラより何和音か、それが問題だ…[カメラ,問題]

上記の例はページを閲覧することにより，ページの内容を元にネタとして記述した Funny のコメントであることがわかるが，無理や問題など特徴語のみの分類では Negative と分類されてしまった．

4.2 構築した分類器によるツイートの分類実験

次に 4.1 節で構築した分類器を用いて分類を行い，人手によりラベル付けしたツイートを正解データとして評価を行った．

4.2.1 ツイートの取得

本研究で用いるツイートを取得するため，ツイートをリアルタイムで取得できる Twitter Streaming API[14] の statuses/sample を利用し，2014 年 5 月 26 日～6 月 3 日に投稿された日本語のツイートのみを収集し，表 2 の条件を満たすツイートを抽出した．収集したツイートの規模を表 7 に示す．

表 7　収集したツイートの規模

全ツイート数	5,413,303
URL 付きツイート数	799,501
表 2 の条件に一致するツイート数	53,884

また，分類器の精度を確認するため，抽出されたツイートを無作為に選択し，人手によるラベル付けを行った．クラス別のラベル付け数を表 8 に示す．

表 8　人手によるツイートのラベル付け結果

クラス	ツイート数
Useful	32
Funny	38
Negative	43

4.2.2 ツイートの分類結果

表 8 のデータを正解データとし，人手によりタイトルと意見部分の分離を行い，4.1 節で構築した分類器でツイートの印象の分類を行った． 3.2 節で述べた顔文字と笑うを意味するネットスラングの有無で比較を行い，各条件における 4 クラスの F 値を表 9 に示す．表の「顔文字」列は顔文字の有無を示し，「www」列は笑うを意味するネットスラングの有無を示している．各クラスの分類結果において最も F 値が高い結果を太字で示している．

表 9　ツイートの分類結果

顔文字	www	Negative	Positive	Funny	Useful
なし	なし	58.82	64.65	46.73	44.00
なし	あり	**65.71**	76.92	65.52	47.83
あり	なし	64.52	**77.36**	48.98	47.06
あり	あり	**65.71**	**77.36**	**67.80**	**51.06**

表 9 より，全てのクラスの分類において，顔文字と笑うを意味するネットスラングを含めた場合の分類精度が最も高いことがわかった．最も F 値が高い条件である顔文字とネットスラングを含めて分類を行った場合の再現率，適合率，F 値を表 10 に示す．

表 10　ツイートの分類精度最大時の結果

クラス	Negative	Positive	Funny	Useful
再現率	69.70	74.55	76.92	41.33
適合率	62.16	80.39	60.51	66.67
F 値	65.71	77.36	67.80	51.06

表 10 より，Positive, Negative の分類精度は高いが，Useful の分類精度が低いことがわかった．正解は Useful であるが分類器で他のクラスに分類してしまう結果が多く，他のクラスと比較し，特に再現率が低いことが課題である．

Useful について分類に失敗したデータを確認したところ，はてなブックマークのそのクラスに分類されるコメントと比較し，話し言葉や顔文字などが用いられており，誤って他のクラスに分類してしまうことがあった．また，インターネット特有の表現の解析に失敗したり，特徴語が選択さ

れないため分類に失敗しているものも見られた．例を以下に示す．[]内に示す語が抽出された特徴語である．

- う、歌ってしまう！これは。。。　[卯]
- こんなの出たら、即買いやわ。[の,やわ(柔)だ]
- きゃわたん [和, 疼]
- この本いいね！[言い値]

上記の例は，話し言葉であったり，省略した記述のため形態素解析に失敗したと考えられる．

収集したデータから各サービスの使われ方を分析したところ，はてなブックマークは主に自分が役に立つ情報へのメモ代わりとして使われることが多く，Twitter は面白い情報を閲覧者にも紹介するために使われることが多いことがわかった．

例としてあるフォントを紹介するページについて特徴語を抽出し比較を行うと，はてなブックマークでは「フォント」，「使える」，「良い」，「便利」，「メモ」，「まとめ」といった自分へのメモ代わりに用いられているが，Twitter では「メモ」，「使える」だけではなく「かわいい」「エエな」といった面白い印象を持ったことや話し言葉でツイートする例があった．

4.2.3 ツイートからのコメント抽出についての検討

4.2.2 の実験では，ツイートから URL とページのタイトルを除去し，意見を抽出する作業は人手により行った．実際に本分類器を用いて自動的に印象を分類する場合，ユーザの意見部分のみを抽出する処理が必要である．

2014 年 5 月 26 日～6 月 3 日に収集されたツイートの中から，表 8 のツイートとは独立して，表 2 の条件を満たすツイートを無作為に 50 件選択し，人手によりタイトルと意見の分離が可能であるか確認を行った．確認した結果の内訳を表 11 に示す．

表 11　抽出した 50 ツイートの内訳

分類	件数
タイトル＋コメント	17
タイトルのみ	15
アクセス依頼	7
診断・占い・採点	7
ツイートボタンデフォルト	4

HTML ファイルの title タグにおいて，「メニュー ｜ ○×料理店」のようにページのタイトルだけではなく，Web サイトのタイトルも記述されている場合がある．title タグの内容の完全一致または部分一致であり，コメントが分離できるものは「タイトル＋コメント」とし，タイトルのみでコメントが記述されていないものは「タイトルのみ」とした．例として同一の URL に関するいくつかのツイートを示す．

- △□が食べたい！ メニュー ｜ ○×料理店
  →コメント「△□が食べたい！」へ分離可能
- △□が食べたい！ ｜○×料理店
  →タイトルの部分一致のため上と同様に分離可能
- メニュー ｜ ○×料理店
  →コメントが記述されていないためタイトルのみ

ツイートが「更新しました （URL）」などといった，更新通知などページタイトルは含まないが，アクセスを依頼する内容であるものは「アクセス依頼」とした．

ゲームの結果や占いの結果などのツイートは「診断・占い・採点」とした．これらは，URL 付きであるものの，その URL に対するコメントではないため印象を分類する必要はない．

また，Web ページへツイートボタンを埋め込む場合，title タグの内容を使うオプションの他，ページ作成者が指定する任意の文字列に変更できるオプションもある．このように，ツイートボタンの埋め込みタグの解析を行うことにより分離できる可能性があるものは「ツイートボタンデフォルト」とした．この 4 件の中には文字数を削減するため，タイトルをツイートボタン向けに要約したり，ブログでは title タグには記事のタイトルとブログ全体のタイトルが記述されているが，ツイートボタンとしてはブログの投稿者名と記事のタイトルをツイートする設定であるブログもあった．

本結果から，実際にユーザのコメントが含まれていたツイートは 17 件であり，タイトル＋コメントとタイトルのみに関しては，リンク先 URL の HTML ファイルを解析することにより，タイトルの分離が可能であると考えられる．ツイートボタンデフォルトはツイートボタン埋め込みタグの解析によりタイトルの分離が可能だと考えられる．しかし，アクセス依頼と診断・占い・採点に関しては詳細な分析を行い，除去する手法を検討する必要がある．

本実験ではツイートからタイトルとコメントを人手により分離し，コメントに対して分類を行ったが，この処理を行わない場合の結果と比較し，タイトルを除去することによる効果の検証を行った．ツイートの分類精度が最も優れていた表 10 の実験条件において，取得したツイートからタイトルの除去を行わず，タイトルと印象が含まれる状態で分類を行った結果を表 12 に示す．

表 12　タイトルを除去しない場合の分類結果

クラス	Negative	Positive	Funny	Useful
再現率	82.05	57.81	51.61	33.33
適合率	54.24	84.09	61.54	61.11
F 値	65.31	68.52	56.14	43.14

表 12 より，意見部分の分離を行った表 10 と比較し，Negative の再現率は大幅に向上しており，それに伴い Positive と Funny の適合率がわずかに向上している．しかし，他の指標においては分類精度の低下の影響が大きく，F 値においては全てのクラスで低下している．

再現率が向上した Negative について結果を確認したところ，意見部分のみでの分類では誤判定があったものの，ページのタイトル部分に「逮捕」や「不正」といった特徴語が含まれており，分類結果が改善しているものがみられた．

結果より，ツイートの意見部分だけを抽出することにより分類精度が改善されることがわかったが，ツイートには SBM のコメントと比較し，ページタイトルが含まれるものが非常に多く，コメントがないものはタイトルを用いて分類するなど手法を検討する必要がある．

また，今回は分類から除外したリツイートも SBM にはない Twitter の特徴であるため，リツイートも評価に用いるなど，より Twitter に最適化した手法を検討する必要がある．

5. おわりに

ソーシャルブックマークのデータを学習データとして SVM を用いて分類器を構築し，ツイートの分類を行った．インターネット上に書き込まれたコメントであることを考慮し，単純に形態素解析の結果だけでなく，顔文字や笑うを意味するネットスラングを用いることにより，分類精度を改善することができた．

SBM のコメントの分類結果に関しては Funny 以外のクラスでは分類精度を示す F 値が 7〜8 割あるが，5 割程度の Funny の分類精度の改善が課題である．ツイートに関しては F 値が 6 割程度ある他のクラスと比較し，4 割程度の Useful の分類精度の改善が課題である．

今後は，誤って分類された項目について検討を行い，SBM のコメントの分類においては Funny，ツイートの分類においては Useful の分類精度向上を目指す．

また，現在 URL 付きのツイートの本文からユーザの意見部分とタイトルなどの分離は人手で行っているが，リンク先 URL の HTML ファイルを取得しタイトルを解析するなど，意見部分の分離の自動化を目指す．Twitter 特有の行動であるリツイートなどについても手法を検討する．

さらに，顔文字やネットスラングをはじめとするインターネット特有の表現に対する分類精度の改善のため，形態素解析器 Juman の辞書の拡張も行う．具体的には，2011 年時点となっている Wikipedia のタイトル辞書の更新，顔文字の辞書の件数追加，また，Wikipedia と比較し流行に関する語が含まれているはてなキーワード[15]を用いた辞書の追加を行う予定である．

謝辞

本研究の一部は，平成 26 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] Twitter, https://twitter.com/
[2] 総務省，平成 24 年版情報通信白書, http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h24/html/nc123220.html
[3] はてな メディアガイド(2014 年 1-3 月版), https://hatenasales.g.hatena.ne.jp/
[4] はてなブックマーク, http://b.hatena.ne.jp/
[5] Scott A. Golder, Bernardo A. Huberman. "Usage Patterns of Collaborative Tagging Systems." , Journal of Information Science, 32(2). pp.198-208, 2006.
[6] Shilad Sen, Jesse Vig, John Riedl, "Tagommenders: connecting users to items through tags", WWW '09, pp. 671-680, 2009.
[7] 丹羽 智史, 土肥 拓生, 本位田 真一, "Folksonomy マイニングに基づく Web ページ推薦システム", 情報処理学会論文誌, 47(5), pp.1382-1392, 2006.
[8] 山家 雄介，中村 聡史，アダム ヤトフト，田中 克己, "ソーシャルブックマークの特性分析とそれに基づく Web 検索結果の再ランキング手法", 情報処理学科論文誌データベース, Vol.1 No.1, pp.88-100, 2008.
[9] Alec Go, Richa Bhayani, Lei Huang, "Twitter sentiment classification using distant supervision", CS224N Project Report, Stanford, pp.1-12, 2009.
[10] 齋藤 準樹, 湯川 高志, "ソーシャルブックマークを基にした Twitter ユーザの興味語抽出・推薦手法の提案と評価", 2011-IFAT-102(2), pp.1-8 2011.
[11] LIBSVM, http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[12] Ting-Fan Wu, Chic-Jen Lin, "Probability Estimates for Multi-class Classification by Pairwise Coupling.", Journal of Machine Learning Research 5, pp.975-1005, 2004.
[13] Juman, http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[14] Twitter Streaming APIs, https://dev.twitter.com/docs/api/streaming
[15] はてなキーワード, http://d.hatena.ne.jp/keyword/