Vol.2015-DBS-162 No.9
2015/11/26

ソーシャルタグを用いたツイートの印象の推定

岡村 康行[1]　湯本 高行[1]　新居 学[1]　上浦 尚武[1]

概要：本研究ではソーシャルネットワーキングサービス（SNS）に着目し，ページに言及しているツイートの印象を，半自動的に収集した学習データを用いて Support Vector Machine により推定する．SVM では大規模な学習データが必要であるが，人手でのラベル付けは多大な労力を要する．そこで，本研究では SNS で利用されているタグをソーシャルタグとし，ソーシャルタグを用いて学習データを収集する．ソーシャルタグとして Twitter のハッシュタグとソーシャルブックマークサービス（SBM）のタグについて検討を行った．その結果，印象を表しているタグが頻繁に使われている SBM のタグを用いることにする．SBM のコメントに対し，タグを基に教師信号を付与したデータを学習データとして分類器を構築し，ツイートの印象の推定を行う．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，さらに娯楽目的かそれ以外かに分類し計 3 種類のクラスに分類する．

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．近年ではソーシャルネットワーキングサービス（SNS）の普及により，誰でも情報を発信することが容易になっている．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 2,070 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印象が記述されているツイートに着目する．しかし，ツイートを用いて機械学習を行う場合，人手でのラベル付けは多大な労力を要する．また，多くのツイートは Web ページの印象を述べていないことからツイートの選別も課題となる．

一方，ソーシャルブックマークサービス（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，2015 年 6 月時点で 438 万人の会員が登録 [3] している日本最大規模の SBM 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，ソーシャルタグを用いて学習データを収集し 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，ソーシャル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．本研究ではソーシャルタグのうち Twitter のハッシュタグ，SBM のタグについて検討を行う．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，さらに娯楽目的かそれ以外の興味や関心が高い情報に分類し計 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．

$$\boldsymbol{b} = (u, r, t, c) \tag{1}$$

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．

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

[1] 兵庫県立大学
University of Hyogo

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物か （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させた新カテゴリー
(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，あとで買う）

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興味関心の高い情報か，娯楽目的の情報かというクラスに分類する．

学習データを自動的に集める研究もいくつか行われており，Pak ら [10] の研究では顔文字を用いており，Kouloumpis ら [11] の研究では Twitter のハッシュタグを用いている．しかし，日本では日本語入力プログラム (IME) の利用が多く，ASCII 文字がほとんどの英語の顔文字と比較し，顔文字の種類が多く分類が困難である．また，ハッシュタグについては，それぞれの国や言語で利用目的が異なると考えられるため，日本語への応用の可能性については 3.1 節で検討する．

SBM のタグを基に学習し，Twitter を対象に評価を行う研究には，齋藤ら [12] 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，目的が興味や属性を表す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[13] を用い，カーネル関数として Radial Basis Function（以下 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 [14] 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．

我々が 2014 年に発表した研究 [15] において，ソーシャルタグとしてはてなブックマークのタグを基に教師信号を付与したデータを学習データとして分類器を構築し，ツイートの印象の分類が可能であることを示した．本研究では，ハッシュタグについての有用性を検証し，インターネット特有のくだけた語に対する形態素解析の精度を向上させ，分類のクラスを実際に多くツイートされているクラスで定義することにより分類精度の改善を行っている．

3. ソーシャルタグについての予備調査

本研究では，ソーシャルタグとして Twitter のハッシュタグと SBM のタグを対象とする．本研究で着目する感情表現に関連するタグが存在するかを調査する．

3.1 ハッシュタグ

Twitter のハッシュタグを用いて学習データを収集する手法として，2 章で Kouloumpis ら [11] の研究があると述べた．この研究は英語のツイートを対象としており，日本語のツイートにおいて印象を表すハッシュタグが存在するかを調査する．

Kouloumpis らの研究で述べられている英語のツイートで頻繁に用いられているポジティブ・ネガティブを示すタグを表 1 に示す．

表 1 英語の Twitter で頻繁に使用される感情表現のタグ

Positive	#iloveitwhen, #thingsilike, #bestfeeling, #bestfeelingever, #omgthatssotrue, #imthankfulfor, #thingsilove, #success
Negative	#fail, #epicfail, #nevertrust, #worst, #worse, #worstlies, #imtiredof, #itsnotokay, #worstfeeling, #notcute, #somethingaintright, #somethingsnotright, #ihate

次に，日本語のツイートにおけるハッシュタグの利用状況を調査する．Twitter Streaming API[16] の statuses/sample を用いて，日本語のリプライ，リツイート以外のツイートを収集する．

ツイートはユーザが投稿したものではない，機械的に投稿されたものも多く存在する．本研究ではユーザによって投稿された URL 付きのツイートを分類の対象とするためこのようなツイートは除外する．

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名より判断できるため，Twitter 公式クライアント及び Twitter で頻繁に使用されているクライアントのうち PC や携帯端末から投稿するクライアントのみ選択する．選択したクライアントについては末尾の付録 A.1 で述べる．

診断系サービスの診断結果や商品の宣伝など，上記の条件を全て満たしていても，印象を表す語が含まれるがページに関する印象ではないツイートが存在する．頻繁にツイートされている URL については評価に影響を与えるため除外する．具体的な条件については付録 A.1 で述べる．

この条件を満たす，2015 年 4 月 1 日から 6 月 26 日に投

稿されたツイートで頻繁に用いられているハッシュタグを調査した．使用頻度の高い上位 30 個のタグを表 2 に示す．

表 2 Twitter で多く使用されているハッシュタグ

RT した人全員フォローする, 拡散希望, モンスト, 刀剣乱舞, 相互希望, pixiv, agqr, 艦これ, わーーーージャニオタさんと繋がるお時間がまいりましたなのでいっぱい繋がりましょ, とうらぶ, パズドラ, わーーーージャニオタさんと繋がるお時間がまいりましたなのでいっぱい繋がりましょ rt してくれた方で気になった方お迎えです, RT した人にやる, lovelive, ふぁぼしたひとにやる, ラブライブ, ラブライバーは RT, nhk, 祝ってくれる人 RT, OneDirection, 西木野真姫生誕祭 2015, 東條希生誕祭 2015, jr 担同士仲良くなろうぜってことで東西関係なく気になった方フォローする, 相互フォロー, Directioners, TheyreTheOne, マルチバースト, RT した nr さん bot さんフォローする, nowplaying, SoFantastic

頻繁に用いられている上位 1,000 件を閲覧した結果，Kouloumpis らの研究に示したようなタグは日本語にも存在するがほとんど利用されておらず，多くが Golder らの研究で述べられている役割の (1)，(7)，そして共通の関心を持つユーザと知り合うためのタグであった．

この結果より，Golder らの研究の役割の (5) にあたるタグが少ない日本語の Twitter では，感情表現を表すタグとしてハッシュタグを利用することは効果的ではないと判断した．

3.2 ソーシャルブックマーク

SBM のタグを用いた学習データのラベル付けの検討のため，印象を表すタグの使用頻度を確認した．はてなブックマークで人気のブックマークを表すホットエントリーを調査し，使用頻度の高い上位 30 個のタグを表 3 に示す．

表 3 はてなブックマークで多く使用されているタグ

あとで読む, まとめ, 2ch, ネタ, web サービス, iphone, web デザイン, web, web 制作, 仕事, lifehack, 生活, 考え方, twitter, 社会, google, これはすごい, mac, javascript, プログラミング, デザイン, お役立ち, ビジネス, 英語, jquery, 人生, html5, wordpress, css, これはひどい

表 3 において下線を引いてあるタグは極性を表すタグである．「これはすごい」,「これはひどい」といったユーザの印象を直接表すタグと，面白いという印象を表すときに付与する「ネタ」というタグがあることがわかる．

このように，SBM においては感情表現を表すタグが頻繁に使用されていることから，本研究ではソーシャルタグとして SBM のタグを用いることにする．

4. ソーシャルタグを用いた分類器の構築手法

4.1 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，娯楽目的の情報とそれ以外の興味や関心が高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，一般的に娯楽情報を除いた興味や関心が高い情報が求められている場合が多い．従って，娯楽情報を除外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娯楽情報を提示することがよいと考えられる．ユーザの検索の目的によって，提示する情報を変更することにより，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．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
  - Funny：ジョークなどの娯楽情報
  - Interesting：娯楽目的以外の興味や関心が高い情報
- Negative：ネガティブな情報

分類のアルゴリズムを表 4 に示す．

表 4 分類のアルゴリズム

Input : $M_{pos}, M_{neg}, M_{fun}, D$
Output : C
$P_{pos} \leftarrow$ predict (M_{pos}, D)
$P_{neg} \leftarrow$ predict (M_{neg}, D)
if $P_{pos} > 0.5$ AND $P_{neg} < 0.5$ **then**
  $P_{fun} \leftarrow$ predict (M_{fun}, D)
  if $P_{fun} > 0.5$ **then**
    $C \leftarrow$ funny
  else
    $C \leftarrow$ interesting
  end if
else if $P_{neg} > 0.5$ AND $P_{pos} < 0.5$ **then**
  $C \leftarrow$ negative
else
  $C \leftarrow$ other
end if
return C

$M_{pos}, M_{neg}, M_{fun}$ 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 4.4 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．$P_{pos}, P_{neg}, P_{fun}$ はそれぞれのクラスの推定確率である．

まず，P_{pos} と P_{neg} を算出し，どちらかが 0.5 以上であれば分類結果のクラス C は Positive または Negative とす

る．Positive の場合，P_{fun} を算出し 0.5 以上であれば C は Funny であり，0.5 未満であれば Interesting とする．いずれにも該当しない場合は Other とする．

4.2 コメントを用いた特徴ベクトルの作成

SBM 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，形態素解析を行う．形態素解析では形態素に分割し，それぞれに対し，読み，品詞などの情報が得られる．

本手法では，コメント中に出現する印象を表す語を特徴語とする．特徴語として，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．

まず，全てのコメントから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．

$$\boldsymbol{v_k} = (w_{k1}, \cdots, w_{kl}, \cdots, w_{kN}) \tag{2}$$

ここで k は k 番目のコメントを表し，l は特徴語のリストの l 番目の語を表す．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 1，使用されていない場合 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．

また，同一の URL に対して同一文のコメントが複数存在する場合，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．

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 5.1.1 項で述べる．

4.3 話し言葉に対応した特徴ベクトルの改良

コメントには話し言葉やインターネット特有の表現などが多く存在する．本研究では，笑うを意味するネットスラング「www」と形態素解析の曖昧性について手法を検討する．

4.3.1 ネットスラング

特に Funny に分類されるコメントにおいて，ネットスラングで笑うという意味を示す「www」が頻繁に用いられている．通常の形態素解析においては未定義語と分類されるため，以下の条件を満たす場合は w の数に関係なく特徴語として「www」も加える．

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

4.3.2 形態素解析の曖昧性を用いた語の除去

形態素解析器 Juman[17] と形態素解析器 Mecab[18] の Unidic 辞書の 2 種類の手法を用いて形態素解析を行う．

Juman は代表表記の情報もあり，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．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．

Mecab は解析に用いる辞書を自由に選択することができ，Unidic はその辞書の一つである．標準の IPA 辞書に比べ語彙が豊富で，さらに代表表記の情報も得られ，他の辞書と比べて細かく分割されることも特徴である．

インターネットに特有のくだけた表現は，形態素解析において誤って分類されることがある．たとえば「この本いいね！」というコメントは Juman では「言い値」と誤った特徴語が得られてしまう．

そこで，2 種類の形態素解析の出力を比較し，2 種類が同じ出力である語はそのまま特徴語として利用するが，異なる結果である場合は語に曖昧性があるとして除外する．しかし，単純に除外すると必要以上に語が除外される可能性があるため，次の条件を満たす語を特徴語として使用する．

- 解析パターンが同じもの
- 形容詞，名詞の品詞の違い
  無理 (だ) のように解析器によって品詞が異なる場合，名詞と助動詞の組み合わせが形容詞と一致する場合は同一の結果とする
- 連続する名詞
  同一の意味でも語句が 2 つの名詞に分割される場合，結合後の表記が一致すれば同一の結果とする
- 顔文字は一方の出現でも使用

この条件を満たさず，除去された語がある場合「不一致フラグ」を付与し，不一致であることを明らかにする．不一致フラグも，4.2 節で示す特徴語の一つとして分類器の入力とする．

4.4 タグを用いたコメントの自動ラベル付け

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 5 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，幅広く特徴語を学習するため，同一の URL において同一条件に一致するコメントは 1 件のみする．

表 5 クラス名と対応するタグ一覧

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
Funny	ネタ	お役立ち, これはひどい

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 3 種類の分類器

$M_{pos}, M_{neg}, M_{fun}$ を構築する．

5. 実験

5.1 節では，実験に使用するデータを収集し，人手によりツイートのラベル付けを行った．5.2 節では，提案手法による学習アルゴリズムで学習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．次に，5.3 節では，学習データと同じ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．さらに，5.4 節では，人手によりラベル付けしたツイートを学習データとし，交差検定を行った．

分類器の評価指標として，再現率，適合率の調和平均である F 値を用いる．このうち，F 値が高いほど分類精度が良いと言える．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

$$Recall = \frac{R_{Q \to Q}}{R_{Q \to *}} \tag{3}$$

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

$$Precision = \frac{R_{Q \to Q}}{R_{* \to Q}} \tag{4}$$

F 値：再現率及び適合率の調和平均

$$F = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \tag{5}$$

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，計算に含まない．

5.1 データセットの構築

実験に使用するデータを収集し，人手によりラベル付けを行う．

5.1.1 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 6 に示す．

表 6 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
名詞・形容詞数	37,437

4.2 節の理由により，多くのコメントで使用されている名詞は除去する．ここでは，我々が 2014 年に発表した研究 [15] より，2,000 件以上のコメントで使用されている名詞は除去する．

4.4 節のラベル付けに用いるタグの利用状況から，各クラスで表 5 に示すタグが多く付与されているページから順に 2,000 件のコメントを学習データとして選択する．

5.1.2 ツイートの取得

本研究で用いるツイートを 3.1 節の手法で取得する．同一の URL に対する同一の文のツイートは除去する．

URL を含むツイートの多くはそのページのタイトルやサイト管理者が指定するデフォルトのツイート文を含んで投稿されている．本研究においてはツイートに含まれる印象部分が重要であり，これらの情報は不要である．また，Twitter は 140 文字の文字制限があることなどを理由に，その URL の HTML の title タグとは異なる場合がある．本研究ではツイートに含まれるページを解析し，HTML の title タグやツイートボタンの埋め込み時に設定できるデフォルトの文字列をツイートから除去する．具体的な除去の対象については末尾の付録 A.2 で述べる．

この中から無作為に選択し人手によりラベル付けを行ったツイートを正解データとして用いる．人手によるラベル付けは 1 ツイートを 2 人で閲覧し，評価が同一である物のみを正解データとして用いる．4 人で 5,000 ツイートを評価して得られたクラス別のラベル付け数を表 7 に示す．また，正解データとして用いることができなかったツイートの内訳を表 8 に，さらに 2 人の評価が不一致であるツイートの内訳を表 9 に示す．「N, F, I」は Negative, Funny, Interesting のいずれかのクラスにラベル付けされたツイートの数である．

表 7 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
Negative	523
Funny	407
Interesting	389
合計	1,319

表 8 正解データとして用いないツイート数

クラス	ツイート数
宣伝目的・スパムツイート	448
印象を表していないツイート	1,355
2 人の評価が不一致	1,878
合計	3,681

5.2 構築した分類器によるツイートの分類実験

4 章の手法で SBM のタグを用いて分類器を構築する．

SVM の RBF カーネルを用い，グリッドサーチによりパラメータを探索した．表 7 のデータを正解データとし，各

表 9　評価が不一致のツイートの内訳

		ツイート数
Negative	Funny	49
Funny	Interesting	167
Interesting	Negative	74
N, F, I	宣伝・印象を表していない	1,588

条件における分類精度の F 値を図 1 に示す．

Juman のみで解析を行い，名詞と形容詞を特徴語として使用した場合を「名詞・形容詞」に，さらに顔文字と笑うを意味するネットスラング www を特徴語に加えたものを「+ネット用語」に，4.3.2 節の曖昧性除去を行いさらに不一致フラグを付与したものを「＋不一致」とする．比較する提案手法は最も分類精度が高い「＋不一致」とする．

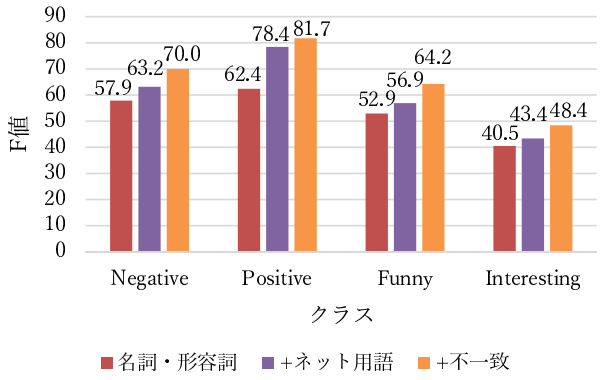

図 1　ツイートの分類結果

図 1 より，Juman のみで解析を行った場合，顔文字と笑うを意味するネットスラングを含めた場合，全てのクラスにおいて分類精度が改善し，Funny については大幅な改善がみられた．

しかし，他のクラスと比較すると Interesting の分類精度が低いことがわかった．分類精度が最大の条件での分割表を表 10 に示す．

表 10　ツイートの分割表

分類器出力		正解データ		
		Interesting	Funny	Negative
	Interesting	168	**128**	79
	Funny	66	287	39
	Negative	85	87	338

42 ツイートは Other に分類されたため除外している．Interesting に分類されたツイートで正解は Funny であるという誤りが多く，適合率の低下の原因となっている．Positive については良い精度で分類できているため，Funny の分類器の改善により精度が向上することが考えられる．

5.3　SBM コメントの分類実験

本手法においては，ラベル付けを行った SBM のコメントにより学習した分類器を用いて，ツイートを分類している．学習データとテストデータの種類の違いが分類精度に与える影響を分析するため，SBM のコメントを対象とし分類する．

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 11 に示す．

表 11　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
Negative	741
Funny	545
Interesting	1,409
合計	2,695

4 章の手法で SBM のタグを用いて分類器を構築する．表 11 のデータを正解データとし，各条件における分類精度の F 値を図 2 に示す．

Juman のみで解析を行い，名詞と形容詞を特徴語として使用した場合を「名詞・形容詞」に，さらに顔文字と笑うを意味するネットスラング www を特徴語に加えたものを「+ネット用語」に，4.3.2 節の曖昧性除去を行いさらに不一致フラグを付与したものを「＋不一致」とする．比較する提案手法は最も分類精度が高い「＋不一致」とする．

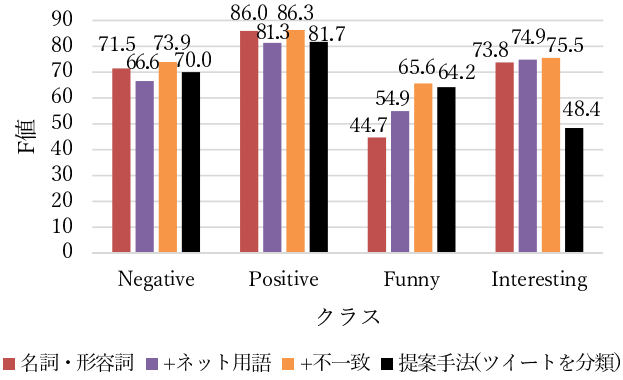

図 2　SBM のコメントの分類結果

図 2 より，5.2 節の実験と比較し，Interesting の分類精度が非常に高いことがわかる．また，Funny は形態素解析の手法を改善することで大幅に分類精度が向上していることが分かるが，5.2 節の実験と比較し分類精度はわずかな差であった．

これははてなブックマークと Twitter で利用目的が異なるためだと考える．3 章の予備調査において，はてなブックマークでは表 3 に示すように，「仕事」「お役立ち」など興味や関心が高い情報に対して付与すると考えられるタグが多く使用されている．一方，Twitter では表 2 に示すよ

うなタグが多く使用されており，娯楽目的で付与するタグが多く見られる．このような利用目的の違いが，提案手法における Interesting の分類精度が向上しない原因だと考えられる．Twitter においては多様な表現が使われており，SBM のコメントを用いて学習した Funny の分類器はこれを満たしていないことが考えられる．

5.4 ツイートで学習したツイートの分類実験

Twitter を対象に分類する場合，予め人手によりラベル付けしたツイートを用いて学習すると，理想的な分類が可能であると考えられる．

そこで，SVM の学習コメントとして，5.1.2 項で人手によりラベル付けを行ったツイートにより学習した分類器 $M_{Tpos}, M_{Tneg}, M_{Tfun}$ を構築する．Positive の学習ツイートは Funny と Interesting とラベル付けされたツイートを使用する．学習データとテストデータで同一のツイートを使用しないよう 10 分割交差検定を行う．学習データ数として 350 ツイート，テストデータとして 35 ツイートを利用し．これらのデータ数は表 7 より Funny の分類器が構築できる上限を考慮して設定した．交差検定の結果を図 3 に示す．

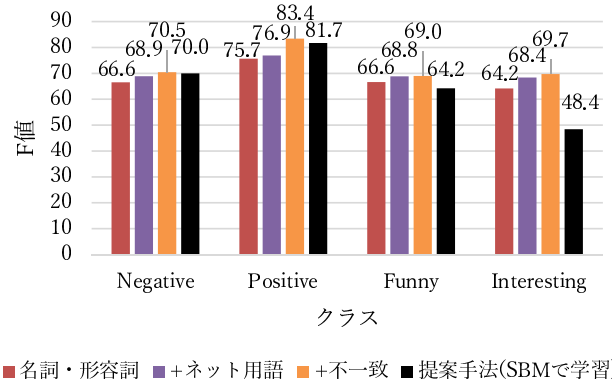

図 3 ツイートで学習した分類器のツイートの分類結果

図 3 より，ツイートに関する分類精度は，5.2 節の実験と比較し，人手によりラベル付けしたツイートによって学習した分類器であれば高い精度で分類できることがわかった．

次に，ツイートの学習データ数の規模の変化による分類精度を確認する．10 分割交差検定における学習データ数を各クラス 100 から 350 ツイートまで変化させたときの分類精度と，比較のため提案手法の分類精度を図 4 に示す．

図 4 より，学習データ数が 350 件においては，提案手法を上回る精度で分類が可能であるが，300 件以下の場合は十分な精度で分類できないことがわかる．また，学習データが少ない場合，特徴ベクトルが生成できないツイートが増加する．特に 100 件以下の場合は特徴ベクトルが生成できないツイートが多く存在した．また，分類器を用いずランダムに分類した場合の F 値と比較したところ，それぞ

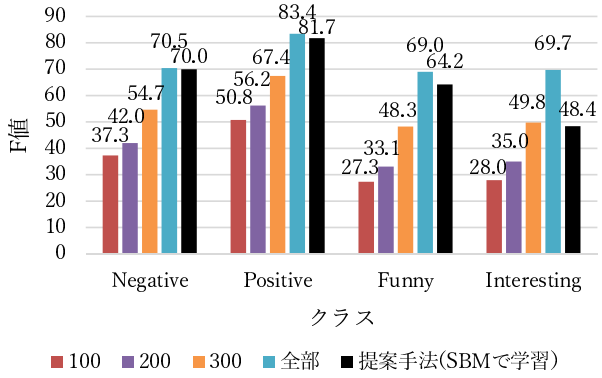

図 4 学習ツイート数を変化させた場合のツイートの分類結果

れのクラスにおいて 100 件における分類精度に近い値となり，現在の評価手法における分類精度の下限であると考えられる．

しかし，人手によるツイートのラベル付けするには多大な労力を要する．ラベル付けしたツイートでの分類は表 7 に示すように，5,000 件をツイートを閲覧しても Negative・Funny・Interesting のいずれかのクラスに分類できたツイートは 25 %程度である．さらにラベル付けしたツイートを実際に学習データとして利用する場合，特徴語が得られなかったり利用できないデータもあるため非常に少なくなる．また，最新のツイートを分類対象とする場合，新しいツイートで利用されている新しいネットスラングや言葉などは学習データに存在しないため，改めて人手によるラベル付けを行う必要がある．

一方，本手法で提案する SBM のタグを用いたラベル付けであれば，予め決まっているタグの選定のみであり，タグを用いて最新の言葉や学習データが容易に取得できるという利点がある．

6. まとめ

ソーシャルタグとして SBM のタグを利用し，教師信号を付与したコメントを学習データとして分類器を構築し，ツイートの印象を Negative・Funny・Interesting のクラスに分類した．機械的に投稿されたツイートやツイートに含まれるタイトルなどを除去することにより，自動的にページの分類が可能である．

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，形態素解析についても複数の解析結果を利用することで分類精度を改善することができた．

構築した分類器でツイートを分類した結果，Negative・Positive については分類精度が高いが，Interesting については改善が必要といえる．

今後は，一つのツイート単位ではなく，ページ単位で集計し，そのページの評判として提示する手法について検討する．

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] Twitter, http://twitter.com/.
[2] 総務省，平成 24 年度情報通信白書，pp.233 http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h24/html/nc123220.html
[3] はてな メディアガイド (2015 年 7-9 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, 2006.
[9] Alec Go, Richa Bhayani, Lei Huang. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, pp.1-12, 2009.
[10] Alexander Pak, Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining Valletta, Malta, pp. 1320-1326, 2010.
[11] Efthymios Kouloumpis, Theresa Wilson, Johanna Moore. Twitter Sentiment Analysis: The Good the Bad and the OMG! Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, pp. 538-541, 2011.
[12] 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.
[13] LIBSVM, http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[14] 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.
[15] 岡村 康行，湯本 高行，新居 学，佐藤 邦弘. 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. 情報科学技術フォーラム講演論文集, Vol.13 No.2, pp.99-104, 2014.
[16] Twitter Streaming APIs, https://dev.twitter.com/docs/api/streaming
[17] Juman, http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[18] MeCab, http://taku910.github.io/mecab/
[19] Tweet Button Twitter Developers, http://dev.twitter.com/docs/tweet-button

付　　録

A.1　ツイートの選択条件

ユーザにより投稿されたツイートを抽出するため，頻繁に使用されている Twitter の公式クライアント及び頻繁に使用される PC や携帯端末から投稿できるクライアントのみ選択する．

URL に多くのユーザが利用している診断系サービスやゲームの URL を含むツイートは除外する．診断の結果やゲームのスコアなどをツイートする際に印象を表す語が含まれるが，これらはページに関する印象ではなく，評価に影響を与えるため除外する．

本文に「レイバン」を含むツイートを除外する．2015 年よりサングラスのブランドの一つであるレイバンの偽ブランド品を宣伝するスパムツイートが多く拡散され，URL も頻繁に変更されている．このツイートは本文に「推奨」「特価」など Positive に分類される語が含まれ，評価に影響を与えるため除外する．

したがって，クライアントや本文や URL により，表 A·1 の条件に一致するツイートを選択する．

表 A·1　本文や URL に関するツイートの選択条件

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 ・TweetDeck・Janetter・twicca・Tweetbot ・Echofon・jigtwi・ついっぷる・Biyon ・Instagram・TheWorld・Mobile Web ・他 21 種類
本文	@で始まり「レイバン」を含まない
URL	URL に関するツイートが複数存在する 下記のドメインではないこと （診断メーカー） http://shindanmaker.com （TryBuzz） http://trybuzz.com （モンスト） http://static.monster-strike.com/ http://www.monster-strike.com/

A.2　タイトル等の除去の概要

ツイートに含まれるタイトルやデフォルトのツイート文を除去する．具体的には下記の 5 種類の手法でページに埋め込まれたタイトルテキストを抽出し，ツイートからタイトルを除去する．部分一致も考慮する．

- HTML の title タグに含まれるタイトル
- Twitter が推奨する埋め込み形式 [19]
  data-text 属性をタイトルとして抽出
- アメーバブログ特有の埋め込み形式
  get の title パラメータを抽出し URL デコード
- Yahoo! ニュース特有の埋め込み形式
  専用の text パラメータを抽出し URL デコード
- Amazon 特有の埋め込み形式
  2 回 URL デコードし text パラメータを抽出