文体の違いを考慮したマイクロブログの極性判定

前田 裕　　嶋田 和孝　　遠藤 勉
九州工業大学大学院 情報工学府
{h_maeda, shimada, endo}@pluto.ai.kyutech.ac.jp

1　はじめに

近年，評判分析の対象として，Twitterが注目されている．Goら [2] は，「:)」や「:(」などのemoticonを利用した訓練データの獲得と機械学習による極性判定（肯定・否定判定）の手法について提案した．Jiangら [3] は，Twitter特有の機能である「リプライ」などに着目し，極性判定する手法について提案している．Brodyら [1] は，Twitterによく見られる「Cooool」のような繰り返し表現に着目し，その繰り返し表現の正規化や感情との関連性などについて検証している．

ここで，我々はTwitter上で用いられる文体に着目する．Twitter上にはさまざまな文体が存在し，その特性が異なる．例えば，話し言葉に近い口語体では，特徴的な文末表現や記号などによる顔文字の多用によってその感情を表現する傾向があるのに対し，書き言葉的な文語体では，言語表現そのものによってその感情が表されることが多い．このような文体の違いを考慮することは，さまざまな場面に有効であると考えられる．

本論文では，このような文体の違いに着目した極性判定の手法を提案する．文体の判別については，人手によるルールと機械学習を併用する．極性判定のための訓練データの獲得については，Goらの手法と同様に，いくつかのemoticonや言語表現を利用する．文体を考慮した手法とそうでない手法を比較し，提案手法の有効性を検証する．

2　文体の定義

文中で利用される語や表現にはいくつかのバリエーションがある．機能語の出現頻度などの文体情報の利用は，blogの性別推定に関するタスク [5] などでその有効性が述べられている．

本稿では，極性判定というタスクの基，書き手の年齢や性別，どのようなネットコミュニティに属しているかでその文体が異なる点について着目する．例えば，「学校」という語の持つ印象は年代によって若干異なると考えられる．実際のTwitterでの使用例を観察すると，若年層の間では「学校の宿題メンドイ …」のような否定的文書が多いのに対して，社会人の間では回想として肯定的に語られることが多いといった傾向の違いが見られる．このような場合，その語の極性は書き手によって異なるため，一つの分類器では適切に扱えない可能性が残る．別の例としては，「＼(ˆ o ˆ)／」という顔文字の肯定性/否定性の解釈がある．この表現は一般に喜びを表す顔文字として使われることが多い．一方で，2ちゃんねるを中心としたネットコミュニティではしばしば否定的表現として用いられる．このような場合も，最終的な文の極性判別の際に誤識別の原因となる．一方で，これらの点を適切に区別して扱うことができれば，極性判定の精度は大きく向上すると考えられる．

以上のような考察から，本稿では，表1のような3種類の文体を導入する．それぞれの分類の基準は，表2に示される主観的な印象に基づいている．「硬い文体」は比較的高い年齢層で使われ，「緩い文体」や「荒い文体」は若年層で使われることが多いため，世代で異なる極性の切り分けに有効である．同様に，コミュニティが異なる「緩い文体」と「荒い文体」とでは，使われる表現も大きく異なっており，文体の切り分けは極性判定に有効に作用すると考えられる．

表 1: 各文体の例

文体	例文
緩い文体	夜桜とってもキレイだったよ〜
荒い文体	夜桜ガチでやばかったわ
硬い文体	夜桜の美しさはかつてないほどだった。

表 2: 各文体の特徴・印象

文体	特徴・印象
緩い文体	口語調，女性的，平和的，緩やか
荒い文体	口語調，男性的，堂々，2ch的
硬い文体	文語調，長文が多い，教科書的

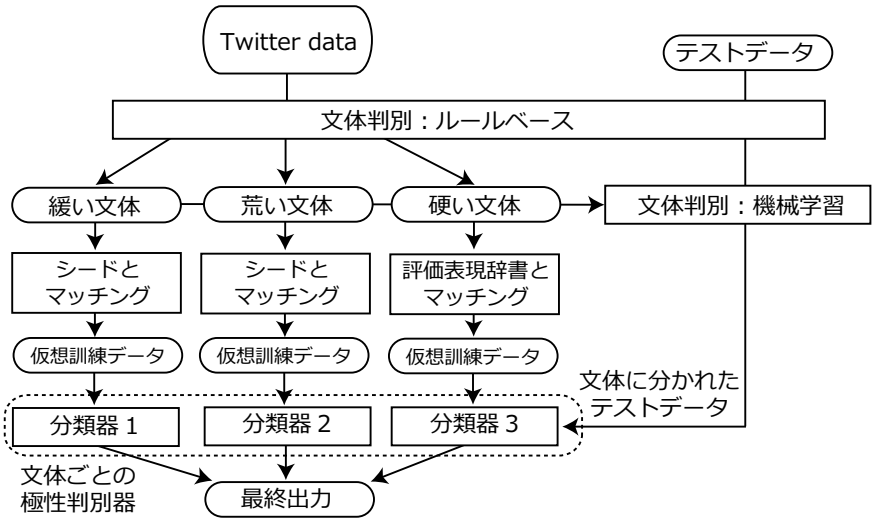

図 1: 提案手法の概要

3 提案手法

本節では，提案手法について述べる．処理の流れを図 1 に示す．提案手法では，まず，入力を前節で定義した 3 つの文体に分類する．その後，シード表現や評価表現辞書などを用いて，仮想的な訓練データを自動的に作成し，文体ごとに分類器を作成する．評価の際には，テストデータを同じような方法で文体ごとに分け，その文体に応じた分類器で極性判定を行う．以降，まず，文体判別の手法について述べ，次に訓練データの作成手法と極性判定のための分類器について述べる．

3.1 文体判別

入力データを文体ごとに分類するために，文体判別を行う必要がある．本論文では，状況に応じてヒューリスティックなルールによる判別と機械学習による判別の 2 つの方法を利用する．それぞれの文体判別方法について具体的な説明を以下に行う．

3.1.1 ルールによる文体判別

まず，ヒューリスティックなルールに基づいて文体を判別する．この手法は，再現率よりも精度を重視した文体判別を行うためにある．ルールとしては，文体ごとに表 3 のようなパターンを数百ほど用意し，正規表現を用い，マッチングによって文体を判別する．文体を判別できる特徴は主に語尾や記号に表れる傾向があり，語尾一致と文中一致の 2 種類を用意した．いずれのパターンにもマッチしない場合には，どの文体にも分類されない．後述する極性判定のための分類器の作成に用いる訓練データを獲得する際には，このルールが適用されたもののみが利用される．

表 3: 文体判別の語尾パターン例

文体名	語尾パターン
緩い文体	だよね，だよぉ，だよ〜，ね！，だもん
荒い文体	だな，だな w，だよな，だろ，かよ
硬い文体	だ。，である。，です。

表 4: 文体判別の文中パターン例

文体名	文中パターン
緩い文体	☆彡，≧∇≦，わーい
荒い文体	ざまぁ，きめえ，お前，俺
硬い文体	(なし)

3.1.2 機械学習による文体判別

ルールによる文体判別は精度重視の枠組みであり，再現率は低くなる．これは，次節で説明するように，高精度な文体ごとの仮想訓練データを得るためである．一方で，テストデータを評価する際に文体を分類する場合，あまりに再現率が低いと，そもそも適切な分類器を適用できない可能性がある．そこで，再現率を重視したアプローチが必要となる．

ここでは，ルールベースで分類された各文体を訓練データとする機械学習を適用する．具体的にはナイーブベイズ分類器を利用し，素性としては文字 3-gram を利用する．実際には，各文体に対して，ルールベースの手法で得られた 10 万ツイートが学習データとして自動的に与えられる．

3.2 訓練データの取得

文体ごとに極性判定を行うためには，機械学習に適用する訓練データが必要となる．しかしながら，十分な訓練データを人手で用意することは高コストである．本論文では,「緩い文体」と「荒い文体」のための訓練

表 5: シード表現の例

文体名	肯定シード	否定シード
緩い文体	♪	;)
荒い文体	きたこれ，おっしゃ	地獄，糞すぎ

データ獲得にはシード表現を用い,「硬い文体」のための訓練データ獲得には既存の評価極性辞書に基づくスコアリング手法を適用することで対応する．

まず，シードを用いた方法について説明する．Turney[4] は，レビュー文書の極性判定において，語やフレーズの極性値を自動的に算出するために，“excellent”や“poor”といった肯定および否定の極性をよく表す語（シード）との共起度を利用する手法を提案した．Go ら [2] は，emoticon を利用して訓練データを自動的に獲得している．本研究でもこれらに倣い，いくつかのシード表現を Twitter のデータに適用することで，仮想的な訓練データを獲得する．具体的には，表 5 に示す表現を，各文体のシード表現として利用する[1]．例えば「緩い文体」に関しては，ルールベースの文体判別手法で得られたツイート群で「♪」を含んでいるものを仮想的な肯定訓練データ,「;)」を含んでいるものを仮想的な否定訓練データとして扱うことになる．なお，各文体において，訓練データ数は肯定事例 10 万ツイート，否定事例 10 万ツイートを取得する．

次に，評価極性辞書を用いた方法について説明する．「硬い文体」の訓練データは，評価極性辞書を利用したスコアリングによって獲得する．本論文では，高村らの単語感情極性対応表 [6] を利用する．この対応表には各単語に対して極性の強さを示す値が 1 から-1 の範囲で付けられいる．獲得処理では，対象となるツイート中に辞書のエントリと一致するものがあれば，その語に対応したスコアを加算する．最終的には，その合計値をそのツイートのスコアとする．本論文では，合計値が 1.8 以上のツイートを肯定事例，-1.8 以下を否定事例として扱った．獲得される訓練データ数は「緩い文体」および「荒い文体」と同様である．

3.3 文体ごとの極性判別

本手法では，前節までに得られた仮想的な訓練データを用いて，各文体ごとに極性判別のための PN 分類器を作成する．すなわち，提案手法では,「緩い文体」のための分類器,「荒い文体」のための分類器,「硬い文体」のための分類器の 3 つが作成される．各分類器はすべてナイーブベイズ分類器であり，素性としては，文字 n-gram と単語 n-gram[2]について最適な組み合わせを用いる．テストデータを評価する際には，ルールベースおよび機械学習に基づく文体判別手法を適用し，同定された文体に対応する分類器へ渡され，最終的な極性判断が行われる．

4 実験

4.1 実験設定

訓練データなどの作成のために，Twitter から収集した約 1 億ツイートを利用した．これとは別に，評価用のテストデータとして，692 ツイートを用意した．この 692 ツイートは人手によって肯定もしくは否定のラベルが付けられている．ここで，全テストデータ中，315 ツイートが肯定文であり，377 ツイートが否定文である．実験では，提案手法が出力した極性との一致率を評価尺度として利用した．

4.2 結果と考察

実験結果を表 6 に示す．文体名の隣にある括弧内の数字は，ルールベースもしくは機械学習に基づく文体判別によってその文体へ分類されたツイート数を表している．例えば,「緩い文体」には 692 ツイート中 307 ツイートが分類されたことを表している．表中の「分類器（緩）」,「分類器（荒）」および「分類器（硬）」は各文体ごとの極性判定用の分類器を意味し，素性はその手法でもっとも精度がよかった場合の素性を表している．表で，例えば,「緩い文体」のための分類器である「分類器（緩）」で,「緩い文体」と判断されたツイートに対する極性判定の精度は 88.3%であり，同様にもし「分類器（緩）」で異なる文体である「荒い文体」のツイートを極性判定した場合の精度は 73.1%であることを表している．すなわち，表中の対角成分（88.3%，82.8%，78.6%）が，ある文体と判定されたツイートを適切な分類器で極性判定した精度を表しており，それ以外の場所は，文体と分類器にミスマッチが起きている状態の精度を表している．各文体で，もっとも精度が高いのは，提案手法が適切に適用されたその対角成分の数値であり，提案手法の有効性が確認された．

表中の「全体」と書かれた行は，ある特定の分類器のみで全データを分類した場合の精度を表している．これはすなわち，従来からある，文体を考慮していない分類器と等価であり，その精度と各文体ごとにカスタマイズされた分類器が正しい文体を分類した場合の精度比較することで，文体を考慮する提案手法の精度

[1] 表中の「;)」は泣いた顔文字や汗をかいた顔文字をターゲットとした意図がある．

[2] 形態素解析には MeCab（http://mecab.sourceforge.net/）を利用．

表 6: 極性判定精度

P/N 分類器	分類器（緩）	分類器（荒）	分類器（硬）
素性	文字 3-gram	文字 4-gram	文字 4-gram
緩い文体 (307)	**88.3**	71.0	76.3
荒い文体 (145)	73.1	**82.8**	77.9
硬い文体 (240)	75.0	72.9	**78.6**
全体 (692)	**80.5**	73.0	77.1

表 7: 極性判定精度 （ルールによって文体判別できた tweet のみを対象とした実験）

P/N 分類器	分類器（緩）	分類器（荒）	分類器（硬）
素性	文字 3-gram	文字 4-gram	単語 1,2,3-gram
緩い文体 (239)	**90.4**	69.9	79.9
荒い文体 (49)	71.4	**87.8**	83.7
硬い文体 (168)	78.0	75.0	**81.0**
全体 (456)	**83.8**	73.7	80.7

向上率が把握できる．具体的には，対角成分の平均値は，83.2%であり，これが提案手法の全体的な精度である．一方で，文体を考慮していない場合の精度は，「分類器（緩）」の場合が最高で 80.5%であり，提案手法の精度の方が 2.7%高いことになる．最低の場合である「分類器（荒）」と比較すると，その精度差は 10.2%（73.0% vs. 83.2%）であり，提案手法の有効性が確認できる．

表 6 の結果から，文体に着目することの有効性が示されている．一方で，この実験では，文体の判別はルールと機械学習によって自動的に行っており，例えば「緩い文体」に分類された 307 ツイートが，本当に「緩い文体」の範疇のツイートである保証は必ずしもない．特に機械学習に基づく文体判別では，文体判別の再現率を重視するために各ツイートを強引にいずれかの文体に割り振っており，正確性の面で大きく劣る．文体の判別ミスは，分類器と文体とのミスマッチと等価であり，極性判別の精度にも大きく影響する．そこで，より本質的な提案手法の精度を検証するために，精度の劣る機械学習による文体判別は行わず，精度の高いルールベースでいずれかの文体に分類されたツイートのみを対象とした実験を行った．この場合，ルールでどれにも当てはまらなかったテストデータ中のツイートは，評価対象から外れることを意味する[3]．その場合の実験結果を 7 に示す．この場合の提案手法の精度は，86.4%（表 7 の対角成分の平均）であり，文体を考慮しない手法と等価な場合の最高精度である 83.8%（分類器（緩）を全てに適用）を上回っている．さらに，テストデータを全て扱った場合の平均精度である 83.2%と比較しても 3%強の精度向上がみられた．この結果は，文体判別手法の精度が極性判別の精度に大きく影響することを裏付けるものであり，今後，より精度の高い極性判別を行うためには，文体判別手法の精度向上が不可欠であることを示している．

素性について着目すると,「緩い文体」や「荒い文体」では文字 n-gram が,「硬い文体」では単語 n-gram が最も有効であった．これは，口語体には文法逸脱，顔文字などが多く，形態素解析のエラーが生じたためであると考えられる．

[3]具体的にはテストデータの数が 692 ツイートから 456 ツイートに減少した．

5 おわりに

本稿では，Twitter を対象とした極性判定において，文体の違いを考慮した手法を提案し，評価した．文体を考慮しない手法と比較して，最高で 10%程度，精度が向上し，提案手法の有効性が確認された．提案手法において，事前の文体判別が最も重要であることも実験により確認された．より高い極性判別精度を実現するためには，より精度の高い文体判別手法が必要であり，今後の課題の一つである．

参考文献

[1] S. Brody, N. Diakopoulos. “Cooooooooooooooollll-lllllllll!!!!!!!!!!!!!! Using Word Lengthening to Detect Sentiment in Microblogs”, In Proceedings of EMNLP 2011.

[2] A. Go, R. Bhayani, L. Huang. “Twitter sentiment classification using distant supervision”, Technical reports., Stanford University 2009.

[3] L. Jiang, M. Yu, M. Zhou, X. Liu, T. Zhao. “Target-dependent Twitter sentiment classification”, In Proceedings of ACL, 2011.

[4] P. D. Turney. “Thumb up? or thumbs down? semantic orientation applied to unsupervised classification of reviews”, In Proceedings of ACL, pp. 417-424, 2002.

[5] 池田, 南野, 奥村. “blog の著者の性別推定”, 言語処理学会第 12 回年次大会, 2006.

[6] 高村, 乾, 奥村. “スピンモデルによる単語の感情極性抽出”, 情報処理学会論文誌, Vol. 47, No. 2, pp. 627-637, 2006.