SVMを用いた顔文字の感情極性推定

奥山恭平　　松本 忠博
岐阜大学大学院工学研究科

{okuyama, tad}@mat.info.gifu-u.ac.jp

1　はじめに

近年、メールやソーシャルネットワーキングサービス（SNS）、商品のレビューなどにテキストだけではなく、顔文字も使うことが多くなってきている。さらに、TwitterやLINEのようなコミュニケーションツールは一言二言といった非常に短いテキストが使われていることが多く、その一言二言も略語や口語といった自然言語処理では難しい処理となる。そのため、テキストのみの感情、または感情極性推定はさらに困難となる。

しかし、こういった短いテキストが使われている場合の多くは短いテキストの後に感情表現として顔文字が使われている。したがって、評判分析など実際のレビューを扱った研究や、SNSを扱う研究には顔文字が肯定的（Positive）であるか、または、否定的（Negative）であるかといった感情極性を考慮することが必要になってくる。

そこで、本研究では実際に使われている顔文字の感情極性を推定するため、Twitterから顔文字を収集し、顔文字に対して感情極性値を割り振り、SVM（サポートベクターマシン）で感情極性の推定を行い、その推定結果を述べる。

なお、本研究では感情極性の分類の際、後ほど3.3で述べる顔文字感情極性対応表を使用する。

2　関連研究

先行研究 [2] では、顔文字を正規表現にて抽出し、その顔文字を単一、連接2記号、連接3記号の3種類に分割し、その記号、または記号の組み合わせを感情別に出現する頻度を求め、その頻度を感情極性毎に加算することで顔文字の感情極性を分類している。

また、文献 [3] では、Word2vecを使用した、事前にラベル付けの必要がない、教師なし学習による顔文字の感情推定を行っている。

本研究では、顔文字の記号一つ一つに感情別出現率を割り振り、それを単語ベクトルのように扱うことによって、SVMを使用した顔文字の感情極性推定を行う。

3　顔文字の感情極性推定

顔文字の感情極性推定は以下の手順で行う。

1. 顔文字のデータを収集を行う。
2. 顔文字のデータに対して、正規表現にて抽出を行う。
3. 顔文字のデータを1記号ずつに分割し、その記号の感情極性別出現率を算出する。この感情極性別出現率をまとめたものを顔文字感情極性対応表と呼ぶことにする。
4. SVMの教師用のデータとして、収集した顔文字のデータを顔文字感情極性対応表に従ってベクトル化し、モデルを作成する。
5. Twitterから評価データとなる顔文字を収集する。
6. 評価データとなる顔文字を顔文字感情極性対応表に従ってベクトル化し、SVMにて感情極性の分類を行う。

3.1　顔文字データの収集

顔文字のデータは顔文字図書館 [1] より収集する。収集した顔文字の中には感情極性を含まないものもあるため、その顔文字を除外する。

顔文字が感情極性を含むか、含まないかの判断に関しては、人手で行う。

その結果、顔文字をポジティブだと判断したものが1050個、ネガティブだと判断したものが1100個となった。

3.2 正規表現による顔文字の抽出

先行研究と同様に、正規表現にて抽出を行う。
本研究では以下の正規表現を使用した。

$$[\hat{}0-9A-Za-zあ-ヶ一-龠]*[vmo\ \mathrm{v\ m}]* \quad (1)$$

$$[(]+ \quad (2)$$

$$((?![0-9A-Za-zあ-ヶ一-龠]\{2,\}).)\{2,\} \quad (3)$$

$$[)]+ \quad (4)$$

$$[\hat{}0-9A-Za-zあ-ヶ一-龠]*[vmo\ \mathrm{v\ m}ノ\ヽ]* \quad (5)$$

この正規表現は、1 が顔の右側、2 が右頬、3 が顔のパーツ、4 が左頬、5 が顔の左側を表している。

この正規表現で抽出を行った結果、抽出できた顔文字は、ポジティブが 965 個、ネガティブが 971 個という結果になった。

3.3 顔文字感情極性対応表の作成

先ほど抽出した顔文字を 1 記号ずつに分割する。その分割した記号に対して、その記号が各感情極性に対して、いくつ出てきているかという出現率をポジティブ側、ネガティブ側の両方で算出し、これを感情極性別出現率と呼ぶことにする。

$$\frac{\text{ポジティブ側で記号が出現した数}}{\text{全体の記号数}} \quad (6)$$

$$\frac{\text{ネガティブ側で記号が出現した数}}{\text{全体の記号数}} \quad (7)$$

$$(6)-(7)=\text{感情極性値} \quad (8)$$

(6) で算出した値から (7) で算出した値を引き、出現率の差を算出する (8)、この差を記号の感情極性値と呼び、この感情極性値を集めたものを顔文字感情極性対応表と呼ぶことにする。

表 1: 顔文字感情極性対応表の一部

記号	感情極性値
(	0.09090909090909
)	0.082820634169428
^	0.814079422382672
*	0.672131147540984
<	-0.452991452991452
>	-0.442307692307692

顔文字感情極性対応表にある記号は 193 種類あり、その記号一つ一つに－ 1 から 1 までの数値が割り振られている。

3.4 SVM に使用する教師用データの作成

正規表現にて抽出した顔文字のデータに対して、先ほど作成した顔文字感情極性対応表に従ってベクトル化する。図 1

このベクトル化した顔文字を集め、svm-train にて教師用モデルを作成する。

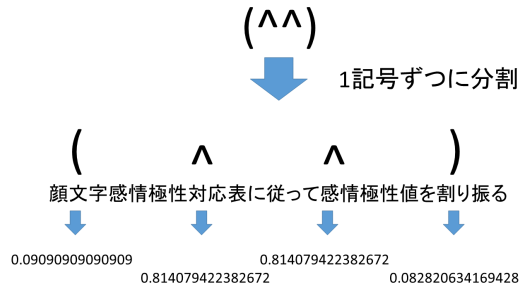

図 1: 顔文字のベクトル化

3.5 評価データの収集

評価データとして、Twitter からの顔文字データを収集する。

収集には、CoreTweet で SearchAPI を用いる。収集する Tweet に関しては、ポジティブ、ネガティブ両方を収集したいので、ランダムに収集している。

収集した Tweet に対して、正規表現にて顔文字の抽出を行い、これを評価データとする。

3.6 評価データの感情極性推定

収集した評価データに対して、感情極性推定対応表に従って、ベクトル化する。

ベクトル化した評価データを 3.4 節で作成した教師用モデルを使い、svm-predict でクラス分け（感情極性推定）を行う。

4 顔文字の感情極性推定実験

前節で述べた手順で顔文字の感情極性推定実験を行った。

評価データには Tweet をランダムに、ポジティブの顔文字が 50 個ネガティブの顔文字が 50 個になるまで収集したものを使用する。

また、抽出出来なかった顔文字、何らかの感情極性を含まないと判断したものに関しては、今回の推定の対象にはなっていない。

本研究では感情極性推定の際に機械学習ライブラリである libsvm-3.21 を使用する。

評価に関しては人手で分類したものを正解とし、それに対してどれだけの精度で分類できたかを計る。

さらに、先行研究の出現率の加算による感情極性分類と比較する。

5　顔文字の感情極性推定実験結果

まず、顔文字の抽出結果について、表 2 のように、括弧がないもの、括弧の片方が欠けているものに関しては抽出出来なかった。

表 2: 評価データに対する顔文字の抽出出来なかった例

顔文字
^_^
(^^v

表 3: 感情極性推定結果

	SVM	加算
ポジティブ	0.76	0.86
ネガティブ	0.84	0.7
全体	0.8	0.78

実験結果を表 3 に示す。左から、SVM による顔文字感情極性の推定結果、加算による顔文字の感情極性推定結果となっている。

顔文字の感情極性の推定結果はポジティブ側では 50 個中 43 個正解という精度で推定できた。「加算」に劣るものの、ネガティブ側では 50 個中 42 個正解という精度で推定できた「SVM」の方が良い結果を得られた。

また全体では SVM による推定の方が 2 %ほど良い結果を得られた。

表 4: 正しく推定できた例

顔文字	推定結果	正解
(v^-)	P	P
(((o(*▽*)o)))	P	P
(〃▽〃)	P	P
(’д’⊂彡☆)) 凵´)	N	N
(つ凵｀)	N	N
(-””-;)	N	N

表 5: 誤って推定された例

顔文字	推定結果	正解	
(´∀｀　*)	N	P	①
(´ -ω-｀) ゞ	N	P	②
ヽ(; ▽ ;)ノ	N	P	③
＼(^o^)／	P	N	④
(^_^;)	P	N	⑤
_(:3」∠)_	P	N	⑥

また、表 4、5 について、左から顔文字、その顔文字の SVM による感情極性推定結果、その顔文字の正解となる感情極性となっており、P はポジティブを、N はネガティブを表している。

表 5 の、誤って推定されてしまった原因について、①と②の顔文字に関しては学習データに「´」と「｀」が同時に含まれている顔文字がネガティブ側に多く、この記号が顔文字の眉として使われているのか、目として使われているのかの判別ができないが為に誤ってしまったと考えられる。これと同様に④と⑤の顔文字に関しても、括弧の中に「^」が 2 つある顔文字はほとんどポジティブであるため、正しく推定出来なかったと考える。③の顔文字に関しては泣きながら笑っているようなポジティブ側の顔文字であるが、泣き顔のようにみえる顔文字は全てネガティブ側に推定されてしまっているのでこの顔文字も同様にネガティブであると推定されてしまっている。

さらに⑥のような顔だけでなく全身が入っているものも正しく推定出来なかった。

6　おわりに

本研究では顔文字感情極性対応表を作成し、SVM を使用した顔文字の感情極性推定を行った。

1 記号ではわずかではあるが SVM による推定の方が上回った。

顔文字感情極性対応表に関しては、顔文字図書館のデータのみを学習データにしているので、今後は実際に Twitter などで使われている顔文字も学習データとして追加していきたい。

顔文字の感情極性推定実験に関しては、ネガティブ側は良い結果を得られたが、ポジティブ側の特に「´」と「｀」が同時に使われている顔文字が多く誤って推定されていたので、学習データの「´」と「｀」が同時に使われている顔文字について見直す必要がある。

また、今後の展望として、この SVM を使った顔文字の感情極性推定をニュートラルを含めた 3 種類に対しても行っていきたい。

参考文献

[1] 顔文字図書館：http://www.kaomoji.com/kao/text/

[2] 三好辰明、太田学、”ツイートに出現する顔文字等の文字と記号に着目した感情分類” DEIM Forum 2013 D9-2

[3] 黒崎優太、高木友博、”Word2Vec を用いた顔文字の感情分類”言語処理学会 第 21 回年次大会 発表論文集 (2015 年 3 月)

[4] 池川知里、新妻弘崇、太田学、”顔文字の役割を利用したツイートの感情極性推定”DEIM Forum 2014 E6-4

[5] CoreTweet:http://coretweet.github.io/docs/index.html

[6] SearchAPI:https://dev.twitter.com/rest/reference/get/search/tweets

[7] LIBSVM:https://www.csie.ntu.edu.tw/ cjlin/libsvm/