G-23

ソーシャルメディアを用いた絵文字の感情極性分類

Sentiment Polarity Classification of Emojis Using Social Medea

木村 真有†　　桂井 麻里衣†
Mayu Kimura　　Marie Katsurai

1. はじめに

絵文字は感情や行動，雰囲気といった非言語的要素を簡単に表現できるため，メールや Twitter，LINE など様々なコミュニケーションサービスに用いられている．そのため，ソーシャルメディアにおける感情分析では，絵文字の表す感情に着目した手法が提案されている[1,2]．具体的に，文献[1,2]では Twitter で使用頻度が高い絵文字を喜び，悲しみ，怒り，嫌悪感の 4 種類に予め手動で分類し，テキスト中に出現する絵文字の感情をそのテキストの表す感情とみなす．しかしながら，文献[1]，[2]のどちらも著者らが独自に絵文字と感情の対応付けを行っており，分析に用いる絵文字の数は少ない．このようなソーシャルメディアの感情分析を促進することを目的とし，大量の絵文字に感情スコアを付与した Emoji Sentiment Ranking と呼ばれる辞書が文献[3]により提案された．文献[3]では，83 人の評価者によって 160 万件のツイートに対しポジティブ，ネガティブまたはニュートラルの感情ラベルを付与し，各感情における絵文字の出現割合を算出した．これにより，ポジティブなツイートに頻出する絵文字ほど高い感情スコア，ネガティブなツイートに頻出する絵文字ほど低い感情スコアが割り当てられる．しかしながら，人手によるツイートの感情ラベル付与には多大な労力が必要という問題点がある．そこで本稿では，このような絵文字の感情辞書を自動で構築するために，感情語との共起頻度に基づく絵文字の感情スコア算出手法を提案する．また，Instagram[1]から取得したデータセットを用いて実験を行い，文献[3]で手動構築された Emoji Sentiment Ranking との相関分析により提案手法の有効性を示す．

2. 提案手法

本章では，ソーシャルメディアにおける感情語との共起に着目した絵文字の感情スコア算出手法を提案する．提案手法では，はじめにデータセット中の各単語に対し，SentiWordNet[2] [4]を用いて感情スコアを割り当てる（2.1）．次に，絵文字と単語の共起頻度に着目した絵文字の感情スコア算出手法を提案する（2.2）．

2.1 SentiWordNet を用いた単語の感情スコア算出

SentiWordNet とは，英語の概念辞書である WordNet[3] 中の synset と呼ばれる同義語グループに対し，ポジティブまたはネガティブの度合いを付与した辞書である[4]．提案手法では，SentiWordNet を用いて単語 i の感情スコア w_iを次式により算出する．

$$w_i = \frac{\sum_{k=1}^{V}(pos_i^k - neg_i^k)}{V_i} \quad (1)$$

上式において，pos_i^k，neg_i^k（$k = 1,2,\dots,V_i$; V_iは単語 i の synset の総数）はそれぞれ単語 i が属する k 番目の synset のポジティブスコアおよびネガティブスコアを表す．算出されるスコアが大きいほど，単語 i はポジティブな感情を表すとみなせる．

2.2 絵文字の感情スコア算出

提案手法では，絵文字の表す感情を自動で分類するために，テキスト内における感情語との共起頻度に着目する．具体的には，2.1 で算出した単語の感情スコアとテキストでの共起頻度に基づき，絵文字 j の感情スコアe_jを次式で算出する．

$$e_j = \frac{\sum_{i=0}^{W} w_i n_{ij}}{\sum_{i=0}^{W} n_{ij}} \quad (2)$$

上式において，W はデータセット中の単語の総数，n_{ij}は単語 i と絵文字 j が同時に出現した回数を示す． 共起頻度のみならず，単語の感情スコアを導入することで，より強い感情を表す単語に重み付けする．最終的に，絵文字の感情スコアが 0 より大きければポジティブな感情を持つ絵文字，0 より小さければネガティブな感情を持つ絵文字として分類する．

3. 実験

3.1 データセット

Instagram では，投稿されたコメントの半数以上に絵文字が含まれていることが報告されている[5]．そこで本実験では，Instagram から取得したコメントを実験用データとして用いる．まず，SentiWordNet におけるポジティブスコアとネガティブスコアの上位単語 300 個をクエリキーワードとしてツイートを収集した．このとき，一人のユーザから取得する投稿は 30 件までとし，1 日以内の連続した投稿は除外した．最終的にデータセット中に含まれるコメントは 643,856 件となった．データセット中の各コメントからは，a, the などのストップワードおよび不要な文字記号を予め削除した．

3.2 絵文字の感情スコア算出結果

実験では，150 件以上のコメントで出現した 266 個の絵文字に対し，式(2)を用いて感情スコアを算出した．スコアが上位 20 個または下位 20 個となった絵文字をそれぞれ表 1，表 2 に示す．266 個の絵文字のうち 264 個がポジティブ，残りの 2 個がネガティブと分類された．負のスコアをとる絵文字は少なく，今回構築した辞書では全体の 0.75%にあたる． 結果から，ソーシャルメディアでは一般にポジティ

†同志社大学，Doshisha University

[1] https://www.instagram.com
[2] http://sentiwordnet.isti.cnr.it
[3] https://wordnet.princeton.edu/wordnet/

表 1. 絵文字の感情スコア上位 20 個

順位	絵文字	感情スコア
1	🎂	0.158
2	💐	0.137
3	🎈	0.136
4	🌷	0.125
5	🎁	0.119
6	🌹	0.119
7	🎆	0.117
8	🎊	0.114
9	🎉	0.109
10	😻	0.108
11	😛	0.107
12	😍	0.106
13	👏	0.106
14	🍒	0.106
15	♥	0.106
16	💟	0.106
17	😘	0.106
18	🌻	0.104
19	💞	0.104
20	🎄	0.104

表 2. 絵文字の感情スコア下位 20 個

順位	絵文字	感情スコア
1	😈	-0.007
2	😷	0.000
3	😪	0.007
4	😠	0.008
5	😡	0.009
6	💊	0.009
7	👎	0.009
8	😵	0.010
9	😤	0.011
10	😖	0.012
11	🔫	0.012
12	🍆	0.013
13	😒	0.013
14	💤	0.013
15	💩	0.013
16	🔊	0.014
17	😰	0.014
18	😫	0.015
19	😞	0.016
20	👹	0.016

ブな文脈で絵文字が用いられると考えられる．この考察は文献[3]で得られた結果と同様である．

次に，文献[3]で構築された Emoji Sentiment Ranking と提案手法により構築した絵文字の感情辞書の相関を分析する．ただし，前者はスコア算出時に用いたツイート数が非常に

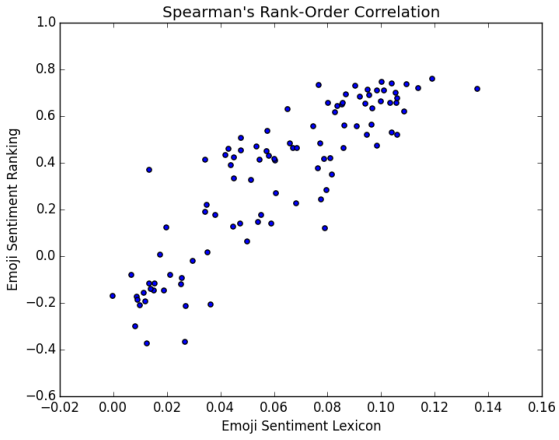

図 1.文献[3]と本研究で提案した辞書の散布図

少ない絵文字も含まれていたため，本実験では文献[3]のデータセットで出現頻度の高い上位 100 個の絵文字のみを分析対象とする．まず，提案手法により算出した絵文字の感情スコアと Emoji Sentiment Ranking [3]に収録されている感情スコアの散布図を図 1 に示す．二つの手法による感情スコアの値域は異なるが，高い相関を示していることがわかる．次に，スピアマンの順位相関係数を算出する．具体的には，文献[3]による絵文字の感情スコアと，本研究で構築した絵文字の感情スコアをそれぞれ降順に並び替え，次式によりスピアマンの順位相関係数を算出する．

$$r_{xy} = \frac{\sum_{j=1}^{n}(x_j-\bar{x})(y_j-\bar{y})}{\sqrt{\sum_{j=1}^{n}(x_j-\bar{x})^2 \sum_{j=1}^{n}(y_j-\bar{y})^2}} \quad (3)$$

上式において，x_j，y_jは提案手法および文献[3]による絵文字 j の感情スコアの順位，$\bar{x}$と$\bar{y}$は各手法による順位の平均順位を表す．r_{xy}の値が 1 に近づけば近づくほど二つのスコア集合の順位相関は強いといえる．算出された順位相関係数は 0.88 であり，高い相関を得られたといえる．

5. まとめ

本稿では，ソーシャルメディアにおける感情語との共起頻度に基づく絵文字の感情スコア算出手法を提案した．Instagram を用いた実験を行った結果， 266 個の絵文字のうち 264 個がポジティブ，残りの 2 個がネガティブと分類された．この結果はソーシャルメディアで絵文字が用いられるのはポジティブな文脈が多いことを示唆しており，文献[3]と同様の考察となった．また，文献[3]の感情辞書との相関係数が 0.88 となったことから，提案した感情スコア算出手法の有効性が示された．

本実験で分析対象とした絵文字は 266 個であるが，今後はデータセットを充足することで，Unicode9.0 以降の絵文字についても分析する予定である．また，構築した絵文字の感情辞書をテキストの感情分析へ応用する予定である．

参考文献

[1] W. A. Hussian, Y. M. Tashtoush, M. Al-Ayyoub and M. N. Al-Kabi, ``Are Emoticons Good Enough to Train Emotion Classifiers of Arabic Tweets?," In *Proc. Int. Conf. Computer Science and Information Technology (CSIT), 2016.*

[2] J. Zhao, L. Dong, J. Wu and K. Xu, ``Moodlens: An Emoticon-Based Sentiment Analysis System for Chinese Tweets", In *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2012, pp. 1528-1531.

[3] P. K. Novak, J. Smailović, B. Sluban and I. Mozetič, ``Sentiment of Emojis," *PloS one* 10.12 (2015): e0144296.

[4] A. Esuli and F. Sebastiani, "SentiWordNet: A Publicly Avail able Lexical Resource for Opinion Mining," In *Proc. Int. Conf. Language Resources and Evaluation (LREC)*, 2006, pp. 417– 422.

[5] T. Dimson, ``Emojineering Part 1: Machine Learning for Emoji Trends." *Instagram Engineering Blog,* http://instagram-engineering.tumblr.com/post/117889701472/emojineering-part-1-machine-learning-for-emoji (Accessed: 08/01/2016)