6Q-01

word2vecを用いた顔文字感情分析手法における クラスタリングの有効性

奥出　悠騎 † 松原　雅文 ‡ Goutam Chakraborty‡ 馬淵　浩司 ‡

岩手県立大学大学院ソフトウェア情報学研究科 † 岩手県立大学ソフトウェア情報学部 ‡

1. はじめに

近年，スマートフォンやインターネットの普及に伴い Twitter，LINE などの SNS によるテキストを用いたメッセージ交換が盛んになっている．SNS は相手の表情や仕草が見えないことから，相手の感情や文に込められた意味など全ての情報を文字列のみから読み取る必要がある．さらに，日本語は抽象的な表現が多いハイコンテクスト文化であるとされ，文字だけでは書き手の意図したものとは別の意味で伝わってしまう可能性ある．

そこで，SNS では自身の感情をより豊かに表現する手段として顔文字が頻繁に使用されている．この，顔文字の感情を分析することで文に込められた感情をより正確に推定することができるものと考えられる．

しかし，SNS の普及に伴い顔文字の種類は豊富になり，現在では約 10 万種類以上が確認されており，その種類は日々増え続けている．そのため，現在確認されている顔文字の全てをリスト化し，把握することは困難である．

本研究では，SNS での顔文字を含むテキストを word2vec で学習しクラスタリングを行うことで，未知の顔文字に対して感情分析を行うことを目的としている．word2vec とは Tomas Mikolv らによって提案されたニューラルネットワークで，分かち書きされたテキストコーパスから単語同士の関係を学習し，単語間の意味をベクトルとして扱うことができる [1]．

本稿では，この word2vec によって算出した顔文字の意味ベクトルを用いて分類実験を行い，実験の結果からクラスタリングの有効性について述べる．

2. 先行研究

word2vec により顔文字を分類する手法として黒崎ら [2] が提案した手法がある．この手法では，6 種類の感情に対してそれぞれ感情を表す語を設定し，感情語と顔文字の類似度を word2vec を用いて算出することで，感情語との類似度が一番高い感情に顔文字を分類している．

しかし，顔文字の中には複数の感情を持つ顔文字や，感情を持たない顔文字も存在するため，顔文字をある特定の感情に完全に分類するには限界がある．

これに対して本手法では，感情を限定せず，顔文字間の類似度をもとにクラスタリングを行うことで，複数の感情をもつ顔文字や感情を持たない顔文字など様々な顔文字に対応することを目指している．

3. 提案手法

3.1. 概要

本手法では，Twitter で収集した顔文字を含むツイートをコーパスとして word2vec で学習を行い，算出した顔文字間の類似度をもとにクラスタリングし顔文字の分類を行う．

word2vec の学習コーパスには Twitter 内の顔文字を含むツイートから助詞・助動詞・固有名詞を除いたものを使用する．

3.2. ツイートの収集

Twitter の公開ストリームから API を用いてツイートを収集する．収集したツイートのうち，リツイート，宣伝ツイート，アスキーアート，記号のみのツイートは学習データとして有効でないと考えられるため除外する．

除外処理を施したツイートに対して，正規表現を用いて顔文字付きのツイートを判別し抽出する．

3.3. 形態素解析

word2vec の学習には分かち書きされたテキストを使用するため，収集したツイートに形態素解析を適用する．ツイートに含まれるユーザ名，URL をあらかじめ取り除き，形態素解析を行い助詞，助動詞，固有名詞，記号を取り除く．最後に,「美味しく」「美味しかっ」といった出現形を「美味しい」のように基本形に統一する処理を行う．

Effectiveness of Emotion Analysis by clustering of Emoticons using word2vec

Yuki OKUDE†, Masafumi MATSUHARA‡, Goutam CHAKRABORTY‡, Hiroshi MABUCHI‡

†Graduate School of Software and Information Science, Iwate Prefectural University, ‡Faculty of Software and Information Science, Iwate Prefectural University

図 1 に実際に正規化した例を示す．図 1 から分かる通り，「@username」を取り除いて形態素解析を行い，「これ / 美味し / そう / ！ / (*’ ω ‘*)」に分かち書きした後，「美味し」を基本形の「美味しい」に統一したものが最終的な学習データとなる．

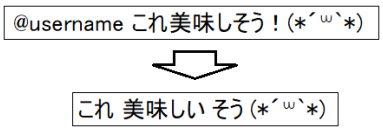

図 1: ツイートの正規化例

3.4. 顔文字のクラスタリング

学習した word2vec から顔文字のベクトルを算出し，顔文字間の類似度を用いてクラスタリングを行う．word2vec を用いて求めた類似度を使うことで，顔文字の意味に基づいた分類を行うことが可能となる．

4. 実験

4.1. 実験概要

顔文字間の類似度を用いたクラスタリングが有効であることを示すために，Twitter で出現した 2,182 件の顔文字に対して K-means を用いてハードクラスタリングによる分類実験を行った．

また，収集したツイートの中で出現数が 100 件以下の顔文字を除いた 1,143 件の顔文字に対して，同様に K-menas を用いて分類実験を行った．

4.2. 実験条件

word2vec の学習コーパスには 2018 年 10 月 1 日から 12 月 1 日までの期間に収集した顔文字を含むツイート 60 万件を使用した．開発言語には Python 3.6，形態素解析器に MeCab，形態素解析用の辞書には mecab-ipadic-Neologed[1] を使用した．

実験結果の評価には純度 (purity) を用いる．式 (1) に純度の計算式を示す．$\Omega = \{\omega_1, \omega_2, \cdots, \omega_k\}$ はクラスタの集合，$C = \{c_1, c_2, \cdots, c_k\}$ はクラスの集合を示す．

$$purity(\Omega, C) = \sum_k \max_j |\omega_k \cap C_j| \qquad (1)$$

4.3. 実験結果と考察

実験結果を表 1 に示す．

表 1: 純度

	K=50	K=100
顔文字 1,143 件	0.75	0.77
顔文字 2,182 件	0.66	0.71

[1] https://github.com/neologd/mecab-ipadic-neologd

表 2: クラスタリング結果の一部

class36	class100
(∩´∀`)∩	(๑˘◡˘๑)
∩^ω^∩	(*´Ⅲ`*)
✌('ω')✌	(*´◡`*)
(*Ü*)	(๑•ω•๑)
٩(๑>∀<๑)۶	(*'ω'*)
ヾ(*'ω')ノ	(*'▽'*)

顔文字 2,182 件を K=100 で分類したとき，純度 0.75 で分類することができた．その際に，喜・哀が複数のクラスタに細分されていたが，その一部を表 2 に示す．class36，class100 共に喜びの感情を表す顔文字だが，class36 が大きく喜んでいるのに対し class100 は喜びと照れが混ざった感情を表している．このため，感情の強弱や用途の違いによって正しく分類されていることが確認できる．

また，出現数が少ない顔文字を減らした 1,143 件 K=50 の実験では 2,182 件 K=100 に比べ純度が上がったため，学習データが少ない顔文字が誤分類の原因になっていることが確認できた．出現数が多い顔文字の中でも一部正しく分類できない顔文字が見られたが，これは複数の感情を持つ顔文字をハードクラスタリングで分類したことが原因として考えられる．

5. おわりに

本稿では，顔文字の web 検索結果およびツイートを用いて学習させた word2vec により顔文字間の類似度を算出し，クラスタリングを行うことで顔文字の感情を分析する手法を提案した．実験の結果から顔文字間の類似度を用いたクラスタリングが顔文字の感情分析に対して有効である可能性が示された．

今後は，顔文字の種類を増やして実験を行うと共に，データ数の少ない顔文字の問題への対策を検討する予定である．

謝辞

本研究の一部は JSPS 科研費 18K11358 の助成を受けたものである．

参考文献

1) Tomas Mikolov, Ilya Sutskever, Kai Chen, GregS.Corrado and Jeff Dean, ”Distributed Represen-tations of Words and Phrases and their Composi-tionality”, Advances in Neural Information Pro-cessing Systems 26 (NIPS 2013).

2) 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 (NLP2015)，B3-3，京都大学吉田キャンパス，March 2015.