卒業論文

ソーシャルメディアの大規模なユーザインタラクションの分析に基づく顔文字分類と予測に関する研究

03-130938　河津裕貴

指導教員　森純一郎 特任講師

2015 年 2 月

東京大学工学部システム創成学科 C コース

概要

Web 環境の成長の中でもソーシャルメディアの普及は著しく, 人々のコミュケーションの基盤となり始めている. 特に Twitter に代表されるようなマイクロブログサービスの形態をとるソーシャルメディアではユーザ間のインタラクションが顕著であることが一つの特徴と言え, ユーザの率直な意見や反応が表出している. この特性に注目しソーシャルメディアから感情分析を用いた評判抽出や意見抽出を試みる研究が多々なされている. また 140 文字の短文という制約のある Twitter に代表されるように, ソーシャルメディアで用いられる短文テキストにおいては感情を分かりやすく伝える手段として短く感情を表現出来る顔文字が頻繁に併用されている. そこで, 顔文字をテキストに付与された感情ラベルと捉え, Twitter のような短文テキストの感情推定を行う研究が近年行われている. 従来研究では使われる顔文字の少ないアルファベットのテキストを対象にし, 感情ラベルとなる顔文字を事前に与えていた. 一方, ソーシャルメディアにおける日本語テキストに見るように, 日本語テキストにはユーザの感情や心境を表現する多種多様な顔文字が利用されている. 本研究の基本的な着想は, これらの膨大な顔文字を分類し, 感情ラベルに資する情報を抽出した上で, テキストの感情推定に利用するというものである.

本研究ではソーシャルメディアにおける日本語テキストの顔文字を対象に, 顔文字の自動分類と感情ラベルに資する情報の抽出, および顔文字を利用した短文テキストの感情推定の手法の確立を目的とする. ソーシャルメディアの代表例として Twitter データを扱い, 研究の前半では大規模なユーザの書き込みデータから多数の顔文字を抽出し, 分散表現モデルとクラスタリング, 及びクラスタの評価関数を用いて顔文字の自動分類を行った. 研究の後半では分類された顔文字をラベルとしてユーザの書き込みから訓練データを作成し, NaiveBayes 分類器を作成して多クラス感情推定及び Positive, Negative, Neutral の感情極性推定を行った. 顔文字分類の結果は従来研究とも親和性のある性質が確認された. またソーシャルメディアにおける大規模データから顔文字を分類する手法を構築し, 本手法による感情推定のベースラインと限界を示した.

キーワード　SNS, 顔文字,Twitter

目次

表目次

図目次

第 1 章

序論

Web 環境の成長は今なお続いており, その中でもソーシャルメディアの普及は著しい. ソーシャルメディアとは個人が誰でも情報発信できるような環境で構築され, 社会的な要素を含み広がっていくよう設計されたメディアの事で, Twitter や Facebook に代表されるような SNS(Social Networking Service) もその一つである. ソーシャルメディアは人々のコミュケーションの基盤となり始めており, 特に Twitter に代表されるようなマイクロブログサービスの形態をとるソーシャルメディアでは個人が気軽に投稿でき, リツイート (他人の投稿を再掲し, 他ユーザに拡散させる機能) やリプライ (投稿に宛先タグを付加し, 返信として投稿する機能) 等のコンテンツに促進され, ユーザ間のインタラクションが顕著であることが一つの特徴と言える. ユーザはたわいもない会話やつぶやきをするだけではなく, 情報の共有やニュースへのコメント, また政治からスポーツに至るまであらゆるテーマについて議論を交わしており, ソーシャルメディア上にはユーザの率直な意見や反応が表出している.

この特性に注目し近年ソーシャルメディアから評判抽出や意見抽出を試みる研究が多々なされている. 例えば, Twitter を利用して, 選挙に関連する書き込みと選挙結果の間に相関があること [1] や, 特定のブランドの商品についてユーザーの評判が抽出できることを示す研究等がある [2]. 評判抽出や意見抽出に際しては, 一般にテキストの特徴から, 書き込みが Positive や Negative であるかの極性の判別や事前に定義された複数の感情やムードの推定が行われる. これらの技術は感情分析と呼ばれ, 自然言語処理技術や機械学習技術を応用したさまざまな手法が提案されている

一方, ソーシャルメディアで用いられる短文テキストにおいては, 文字だけではユーザの持つ微妙なニュアンスや感情の強弱等の非言語情報が伝わり辛いため, テキストにおいて感情を分かりやすく伝える手段が用いられている. 例えば, 近年急速に普及している LINE に代表されるメッセンジャーアプリにおいては, 「スタンプ」と呼ばれる感情や心境を表現したイラストを挿入することで円滑なコミュニケーションが行われている. また, 投稿文が 140 文字の短文テキストという制約のある Twitter においては, 文字や記号列を組み合わせることで作られた様々な種類の顔文字がユーザの感情や心境を表現する手段として頻繁に利用されている.

そこで, 顔文字をテキストに付与された感情ラベルと捉え, Twitter のような短文テキストの感情推定を行う研究が近年行われている [2][3]. これらの手法は, 教師あり学習において顔文

字を正解ラベルとするアプローチである.

従来のアプローチにおいては, 感情ラベルとなる顔文字を事前に与えることにより学習が行われている. 一般にアルファベットのテキストで用いられる顔文字は限定的であり, 例えばアルファベットのテキストの感情分析を行う最近の言語評価タスクにおいては数百の顔文字の極性辞書が定義されている[*1]. 一方, ソーシャルメディアにおける日本語テキストに見るように, 日本語テキストにはユーザの感情や心境を表現する多種多様な顔文字が利用されている. 本研究の基本的な着想は, これらの膨大な顔文字を分類し, 感情ラベルに資する情報を抽出した上で, テキストの感情推定に利用するというものである。

本研究ではソーシャルメディアにおける日本語テキストの顔文字を対象に, 顔文字の自動分類と感情ラベルに資する情報の抽出, および顔文字を利用した短文テキストの感情推定の手法の確立を目的とする. まずソーシャルメディアの代表例として Twitter の大規模データから自然言語処理と分散表現モデル, 及び教師なし学習の技術を用いて多数の顔文字を抽出, 自動分類を行い, 次に分類された顔文字を感情ラベルとして, 教師あり学習である NaiveBayes 分類器によるユーザの書き込みの感情推定を行う. 第 3 章では提案する手法の概要と詳細を記した. 第 4 章では提案する手法を用いた顔文字の分類結果とその考察を行った. 第 5 章では分類結果を用いて感情推定を行い, その結果と考察を行った.

[*1] http://leebecker.com/resources/semeval-2013/

第 2 章

関連研究

まず文化による顔文字の違いや顔文字の分類に関する従来研究を紹介する．次にソーシャルメディアにおける感情分析を, Twitter を例に挙げ紹介する．最後に本研究と同じタスクである顔文字ラベルを用いた感情推定の従来研究を紹介し, 本研究の位置づけを確認する．

2.1 顔文字の分類

2.1.1 東西の顔文字の違い

コンピュータを介した人々のコミュニケーション (CMC:Computer-Mediated Communication) では主にテキストが用いられるが, これは対面型コミュニケーションとは異なり, 話者の持つ微妙なニュアンスや感情の強弱等の非言語情報は文章のみで表現することは困難である．情報伝達の齟齬を防ぐには非言語情報の伝達は不可欠であり, そのためテキストに非言語情報を付加する様々な方法が生まれたが, その代表例が顔文字の存在である [4]．顔文字は主に記号や英数字を用いてテキスト上で人間の顔やジェスチャーを表現するもので, 話者の感情や意図を受け手により分かりやすくする役割を持つ [5]．文字や記号を組み合わせて複数の行で表現されたアスキーアートと呼ばれる記号群も存在するが, 多くの研究では文章中に現れる短い記号群を対象とする．

顔文字の出現時期は議論の余地があるが, 多くの研究では 1980 年代に発生し普及したとされている [6]．CMC の普及が文章中に感情を英数字や記号等を用いて絵的に表現する風潮を導いた．これらの感情を司る象徴的な表現を西洋では *emoticon* と呼ぶ [5]．また顔文字を社会規範とみなし, 文化毎の違いを見ている研究もある [7]．興味深い事に, 地域や文化により使われる顔文字は大きく異なる．例えば東洋人は [(^ ^)] のように, 縦向きに顔として解釈する事ができる顔文字を使う．それに対し西洋人は [:-)] のように, 横向きに解釈できる顔文字を使用している．これは東洋人は感情を人の目から推測するのに対し, 西洋人は口から判断するという違いに起因すると考えられる [8]．また, 顔文字の役割も東西で違いが見られるという研究もある．西洋において顔文字は一種のサインのように文末で用いられるが, 東洋においては実際の会話中の顔を表すように文中の区切りなど様々な位置に出現する [9]．

2.1.2　日本語の顔文字分類

次に日本語の顔文字分類を作成した先行研究を紹介する．田中らは顔文字を形成している記号群に着目した．通常の文章の特徴を単語を素性として表すように, 顔文字を一つの文, 文字を単語と見なして, 文字を素性として扱う事で教師あり学習である SVM による分類を試みた [10]．同じく記号群として扱った研究として Ptaszynski らのデータベース構築がある [11]．Ptaszynski らは顔文字を扱っているサイトから顔文字とカテゴリを抽出し, 感情を判断できない場合は分子運動論をベースに顔文字をパーツ毎に分解し, パーツが登場する他の顔文字の感情から推定した．似た研究として山本らは顔文字の感情を決定するために, 顔文字を口や目等のパーツに分解した [12]．100 個の顔文字に人手でラベリングし, このラベルを用いて分解後の各パーツの感情ラベルを決定した．その後パーツ毎のラベルを参考に新たに 400 個の顔文字に人手で感情ラベルをつけた．アンケート等により感情極性を決定した研究もある．川上らは大学生への調査をもとに, 31 個の各顔文字に対し喜び・哀しさ・怒り・楽しさ・焦り・驚きのそれぞれの感情のスコアを算出した [13]．被験者に対し顔文字に各感情が, 1(全く表れていない) から 5(とてもよく表れている) までの 5 段階で評定させる方法をとっている．江村らは顔文字に貼るラベルの種類にコミュニケーションタイプや動作タイプ等を追加することで, 顔文字推薦の精度が上がった事を確認している [14]．吉田らは Twitter で使われている顔文字に対し, 感情語辞書にマッチングする文中の単語を抽出する事で, 顔文字を事前に用意した 8 つの感情クラスに分類した [15]．

2.2　ソーシャルメディアにおける感情分析

Twitter とは 140 文字以内の「Tweet」と称される短文を投稿できる情報サービスであり, 一般にマイクロブログというカテゴリに分類される巨大ソーシャルメディアである．ユーザ同士のインタラクションが顕著である Twitter については, 多くの感情分析の先行研究が存在する．短文であるという性質や, またソーシャルメディア独特の口語ベースの言葉やネットスラングに起因するくだけた言語などが自然言語処理で正確に分析できない例が多々あり (表 2.1), 感情分析を難しくしている．感情分析には対象とするテキストやタスクにより幾つか種類分け

表 2.1. 非自然言語を含む Tweet の例

えっええーーーーΣ (￣□￣)
おはおー o(^▽^)o
ぁたしの iphone だめぽなた...(´Д‘)
じばにゃん ((’ω’) 三 (’ω’) 三 (’ω’) 三ぐふふ w((’ω’) 三 (’ω’) 三 (’ω’)

される．文書レベルで行うもの [17] やフレーズ毎に行うもの [18], また主観か客観かを判断する分析 [19] 等があるが, 本研究では短文である Tweet を扱うため, 文単位で感情分析する研究

に言及する.

まず感情語辞書のマッチングによる感情分析の手法を Twitter を例に紹介する. 感情語辞書とは, 感情を表す単語を集めたデータベースの総称である. Thelwall らは Twitter 上のテキストは短文なため一般的な感情語を含まない場合が多く, 一般的な文章を対象とする感情語辞書を用いた感情分析では十分な結果が出ないことを示唆している [20]. また感情語辞書とのマッチングによる感情分析の精度を上げるために, 対象とするタスクやトピックに特化した感情語辞書の拡張を行っている [21]. 扱うトピックの文章のみを対象に人手で感情ラベル付けを行い, 文章の感情を利用して含まれる各単語の感情スコアを計算する事で, トピックに特化した感情語辞書を作成している. 同時に彼らは, 感情分析の機械学習を行うにはテキストに人手で感情ラベルを貼る必要があるため作業量が多く困難であり, 精度自体は機械学習による手法の方が良いものの, 感情語マッチングの手法には価値があるとしている. Bravo らはテキストの感情を 1 つ決定するのではなく, 各感情を表す度合いや強弱等を考慮し多次元的な特徴として Tweet の感情を捉えようとした [22]. また顔文字を利用したものとして, Ghiassi らは特定のブランド商品に対するユーザーの評価を Twitter から取得する際, Tweet を感情分析をする際は顔文字を考慮することで大きく精度が向上する事を発見している [2].

2.3　顔文字ラベルを用いた感情推定

この節では顔文字を利用した機械学習による感情推定の先行研究を紹介する. ソーシャルメディアの感情推定において教師あり学習によるアプローチの難点は訓練データの用意である. ソーシャルメディアのテキストには感情ラベルが付いておらず, 1 つ 1 つが短文であるため十分な量の訓練データが必要となる. そのため何かしらの方法で大量のテキストに感情ラベルを貼る必要があるが, 近年感情ラベルとして顔文字を用いる研究が注目を浴びている.

顔文字を利用して訓練データを取得しようという試みは Read らの提唱に始まる [23]. Read らは既存の機械学習による感情推定は訓練用データがトピックや分野に依存しているため汎用的ではない事に言及し, 分野横断的にデータを集める方法として顔文字に着目した. 11 個の顔文字を選出し, それらを含む文章を訓練データとすることで機械学習を行い, 分野非依存の学習器による感情推定のベースラインを示している. Go らは表 2.2 に載せた 8 つの顔文字を含む書き込みを Twitter から取得する事で, 顔文字に則して Positive か Negative かのラベルが付いた訓練データとした [3]. 彼らの実験結果は, 特徴量として Unigram より Bigram を用いた方が精度が良くなる事を示唆している. 顔文字以外の要素も用いる着想として, ハッシュタグに注目する研究が多々ある. Kouloumpis らは Positive と Negative を表す代表的な 2 つの顔文字である :) と :(を用い, さらにハッシュタグも利用して訓練データを構築している [24]. Purver らは顔文字を利用した場合とハッシュタグを利用した場合の感情分析の比較を行っている [25]. Purver らは「happy sad anger fear surprise disgust」の 6 つの感情クラスを用いたが, これに対し Suttles らは心理学者 Plutchik の提唱する wheel of emotions(感情の輪) に基づく 8 つの感情で分類すべきだとし, まず幾つかの顔文字とハッシュタグを 8 つの感情に分類して, その結果をラベルとして訓練データを作成し, 機械学習による感情推定を行った [26].

表 2.2. Go らの研究で用いられている emoticon

Positive	Negative
:)	:(
:-)	:-(
:)	: (
:D	
=)	

本研究は顔文字ラベルを利用した教師あり学習による感情推定を, 大量の顔文字ラベルがある大量データに適用できるよう拡張する事が目的である. 日本の場合は西洋より顔文字の種類が多いため, 少数の顔文字からでは訓練データが十分取得できず, また含む顔文字によりテキストの特性が違うとすれば, データに偏りが生じ不適切だと考えられる. 本研究では, 多量である日本語の顔文字で訓練データを作成するところに違いがある. 前節で述べたように顔文字の分類を行っている研究は多々あるが, いずれの研究も事前に顔文字の区分を人手で作り, 顔文字を割り振る方法をとっている. 用意する感情区分は研究者により様々だが, Twitter のような大規模かつ個人的な会話ベースのインタラクションの中でテキストがどのような感情に分類されるのか主観で決定するのは困難である. 本研究の提案手法では事前に顔文字の区分を決めるのではなく, 顔文字の使われている文脈の情報を用いてクラスタリングし, その後各グループの感情ラベル表現の抽出を行う点に明確な違いがある. また教師なし学習を用いる事で感情ラベル付きの訓練データの必要がなくなり, 大規模なユーザの書き込みデータから取得した良く使われている顔文字を漏れなく分類することが可能である.

第 3 章

提案手法

3.1　提案手法の概要

本研究は顔文字ラベルを利用した教師あり学習による感情推定を, 大量の顔文字ラベルがある大量データに適用できるよう拡張する事が目的である. そのためには事前に顔文字を分類し感情ラベルを貼る必要があるが, 従来の顔文字分類は 1) 人手で顔文字を分類していたため, 扱う顔文字が少数になる, 2) 顔文字の分類区分を事前に主観で決め顔文字を当てはめていたといった問題点がある. そこで本研究では, 大規模なユーザの書き込みデータを用いたクラスタリングによって分類を行う. また顔文字の区分を決めてから分類するのではなく, クラスタの評価関数を用いることで分類数を決定し感情ラベルを検討することで, ユーザの使用傾向に則した分類区分とする. 次に本手法による顔文字分類を用いて NaiveBayes 分類器による感情推定を行い, 本手法を用いた場合のベースラインとなる精度を算出する.

提案手法の構成は 1) クラスタリング, 2) 顔文字のラベル表現の抽出, 3) 顔文字を含む書き込みから顔文字ラベルを基にした訓練データの作成, 4)NaiveBayes 分類器による感情推定 となる (図 3.1).

3.2　顔文字の抽出と分類方法

まず図 3.1 の 1), 2) に当たる, 顔文字の分類とラベル表現の抽出の手法について記す.

3.2.1　顔文字の抽出と分類

本研究では対象とする顔文字をユーザの書き込みデータから取得している. 文章中から顔文字を抽出するアルゴリズムは, 風間らの研究 [27] を用い,

1. 記号関連の Unicode 文字プロパティを持つ文字を探す.
2. 前後に Unicode 文字プロパティを持つか, 日本語以外の文字 (Unicode カテゴリで判定) があるかどうか探索する. ただし間に一つの文字や空白が入ることは許容する.
3. 前後の句読点類を削る (例,「。(^_^)」などがあるため).
4. 事前に与えられた顔文字リストから両側に通常の文字が付いている例外を左右別に抽出

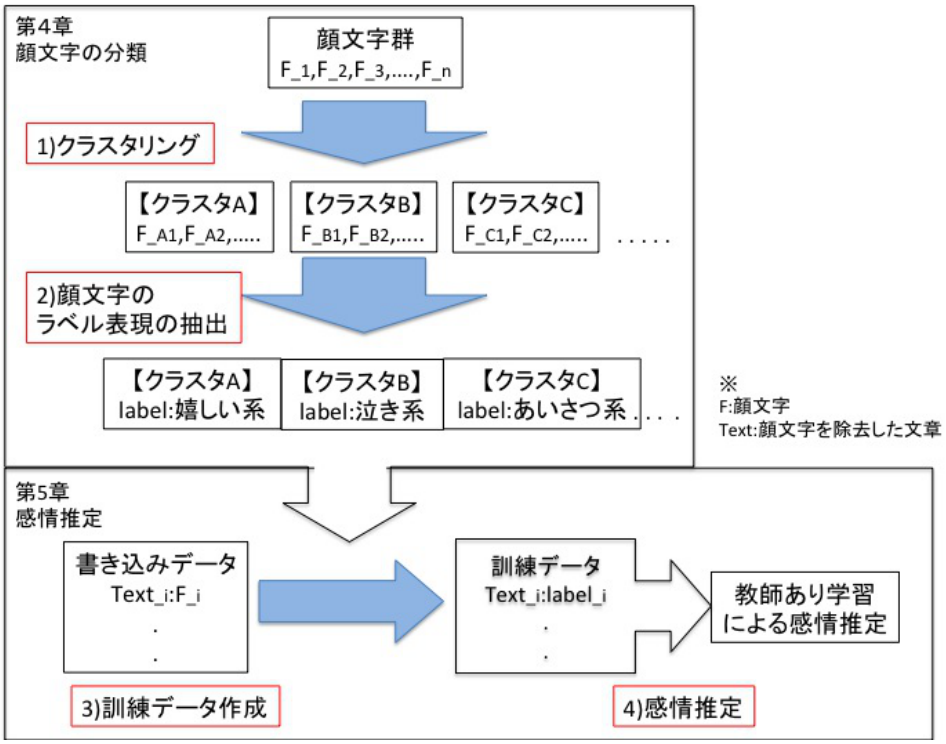

図 3.1. 提案手法の構成

しておき, その例外が存在する場合はその分延長する (例,m(_ _)m の両手の抽出など).

とした.

次に分類の手順は以下のようになる. 顔文字の分類を以下の手順で行う.

1. 1 つ 1 つの顔文字に対し, 分散表現モデルによる単語ベクトルを作成.
2. クラスタリング及びクラスタの評価関数を用いて顔文字を分類.
3. 分類結果の詳細検証によるラベル表現の抽出.

まず分散表現モデルを用いて各顔文字に対応する単語ベクトルを作成する. 分散表現モデルとは, ある単語が用いられる文脈 (コンテクスト) をニューラルネットワークを用いて数百次元のベクトル (これを分散表現と呼ぶ) として表現するモデルのことである. 似た文脈にある単語は似た意味を持つという前提をもとに, 周辺の単語群から注目している単語を推測するようにニューラルネットワークを学習させ, 入力層と中間層の重みを単語ベクトルとして取り出す. 分散表現モデルでは単語の意味だけでなく単語間の関係も表現することもできる. 本研究では単語の分散表現モデルの実装として word2vec を用いる [28]. 作成されるベクトルは, 使用される文脈が類似する単語ほどベクトル同士の距離が短くなる性質を持つ事が知られている. この性質を利用した代表的な例として

vec(東京)+(vec(フランス)-vec(日本)) ≒ vec(パリ)

や

vec(King)+(vec(男性)-vec(女性)) ≒ vec(Queen)

等と計算できることが知られている．本研究では, 各顔文字に対し顔文字の周りの語を抽象化したベクトルを作成することとなる．また word2vec を用いた理由は, 計算時間が短く単語間の意味の距離を考慮しつつ大規模なデータに適応可能なためである．

次に作成したベクトル群を教師なし学習であるクラスタリングを用いて分類する．教師なし学習とは, 入力 (一般に説明変数と呼ぶ) のみの訓練データからモデルを構築するアルゴリズムの事であり, 訓練データに入力と対応すべき出力 (一般に被説明変数やラベルと呼ぶ) が必要ない点で教師あり学習と異なる．またクラスタリングとはデータ解析手法として使われる教師なし学習の一種である．入力データを外的基準なしに幾つかの集合 (クラスタと呼ぶ) に分類する手法で, 入力データには主にベクトルが用いられる．アルゴリズムは階層的クラスタリングと非階層的クラスタリング, また分類対象が複数のクラスタに属せるかどうかによりソフトクラスタリングとハードクラスタリングに分けられるが, 本研究では非階層的かつハードクラスタリングである K-Means クラスタリングを用いる [29]．計算時間を考慮して非階層的クラスタリングとし, また顔文字をラベルとして扱う事が目的なため, 1 つの顔文字が複数のクラスタに属さないハードクラスタリングとした．K-Means クラスタリングはクラスタの中心ベクトルに着目し, 与えられたクラスタ数 K 個に分類する手法である．K-Means クラスタリングのアルゴリズムは以下の通りである．入力データの数を n, クラスタの数を K として,

1. 各データ $x_i(i = 1...n)$ に対してランダムにクラスタを割り振る．

2. 割り振ったデータをもとに各クラスタの中心ベクトル $c_j(j = 1...K)$ を計算する．通常, 割り当てられたデータの各要素の算術平均を中心ベクトルとして使用する．

3. 各 x_i と各 c_j との距離を求め, x_i を最も近い中心ベクトルを持つクラスタに割り当て直す．

4. 2, 3 の処理で全ての x_i のクラスタの割り当てが変化しなかった場合, あるいは変化量が事前に設定した一定の閾値を下回った場合に, 収束したと判断して処理を終了する．そうでない場合は新しく割り振られたクラスタから c_j を再計算して 2, 3 の処理を繰り返す．

となる．

分割するクラスタ数は与える必要があるため, 最適なクラスタ数を求めるために一定の範囲の各クラスタ数でクラスタリングを行い, それぞれに対し以下の式を計算したクラスタの評価関数である Davies Bouldin Index(以下 DB 値) を求める [30]．

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{i\neq j}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right) \tag{3.1}$$

ここで c_i はクラスタ i の中心点, σ_i はクラスタ i の各要素とクラスタ i の中心点との距離の平均, $d(c_i, c_j)$ は中心点 c_i と c_j の距離を表す．$\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}$ の値は任意の二つのクラスタに注目し, 2 つのクラスタの距離が長くかつ各クラスタのまとまりが良いと小さい値となる．DB 値を比較する事で最適なクラスタ数を決定する．

3.2.2　顔文字のラベル表現の抽出

分類された顔文字がどのような傾向を持つのか把握するためのデータとして, 以下の 2 つを抽出する.

1. 分類された顔文字の一覧
2. ユーザの書き込みデータにおいて, 顔文字と共に使用されている単語の一覧

1 では出現頻度の高い顔文字をクラスタ毎に抽出する. 2 では各クラスタごとに特徴語を抽出する. まず顔文字を含む書き込みデータを取得し, 含まれている顔文字の分類に則して書き込みデータを分類し, 顔文字のクラスタ毎に書き込みデータ群を作成する. 各書き込みデータから顔文字を消去したテキストを形態素解析により単語に分解し, 単語の出現回数をクラスタ毎に集計したものを, 共に使用されている単語群とする. 形態素解析とは自然言語処理の基礎技術の一つであり, 事前に作成された文法や単語辞書等を用いて意味を持つ最小単位である形態素毎に分割する分析である. 形態素解析器はオープンソースである MeCab を用いる [31]. 次に単語群から特徴語を抽出する方法として, クラスタ毎に

方法 1. 出現回数の多い単語

方法 2. TFIDF 値の高い単語

方法 3. 総合出現回数に対する各クラスタでの出現割合の高い単語

の 3 種類を比較する.

方法 1 では各クラスタ毎に出現回数の高い単語を抽出する.

方法 2 の手法で使う TFIDF とは文書特有の特徴語を表すスコアである [32]. 文書の特徴を考える際に単純に出現回数の高い単語を取り出すと, 普遍的に出現する一般語が出てきてしまう懸念がある. これらの一般語の重みを下げるために IDF(Inverse Document Frequency) を用いて, 多くの文書に出現する単語の重要度を下げる. また文書長の影響を考慮して, 単語の TF(Term Frequency) も計算しかけ合わせる事で, 各単語の文書における重要度を計算する考え方である. D を総文書数, n_{ij} を単語 t_i の文書 d_j における出現回数, $\{d : t_i \in d\}$ を単語 i を含む文書数として,

$$tfidf_{ij} = tf_{ij} * idf_i \tag{3.2}$$

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{3.3}$$

$$idf_i = log\frac{D}{\{d : t_i \in d\}} \tag{3.4}$$

と計算される. 本研究では IDF の計算の際各クラスタに分類されたテキストを全てつなげたものを 1 文書とする. 例えば 5 クラスタに分類した場合は全部で 5 文書存在する事になる. また IDF 値が高くても TF が極端に少ない単語はノイズとなると考えられるため, 各文書で出現回数が 10 回以下の単語は除去している.

次に方法 3 について, 単語 i の文書 j における登場回数を x_{ij} として,

$$R_j(i) = \frac{x_{ij}}{\sum_j x_{ij}} \tag{3.5}$$

で表される指標値 $R_j(i)$ を用いる. この指標値を用いる理由は, 今回のように大規模なユーザの書き込みデータを少数の文書に集約した場合, TFIDF を用いた特徴語の抽出に際し IDF による重みが上手く機能しない事が懸念されたからである. 方法 2 と同様にノイズを除去するため, 全文書における単語出現回数が 100 回以下の単語は除去している.

3.3 顔文字予測による感情推定方法

次に図 3.1 の 3), 4) に当たる, 顔文字ラベルを用いた感情推定の手法について記す.

3.3.1 訓練データの作成

顔文字を含むユーザの書き込みデータの集合について, 各データから顔文字と顔文字を除いたテキストのペアを抽出する. ここで, 顔文字の現れる箇所がテキストの区切りになっていると仮定し, 文頭から最初に現れる顔文字の直前までをテキストとして扱い, 顔文字以降のテキストは除去する. 抽出した顔文字に対応するクラスタをテキスト感情推定の正解ラベルとし, 残りのテキストから感情推定に用いる特徴量を生成する. 特徴量の生成のため, テキストを形態素解析により形態素に分割し, それらを Bag-of-Words モデルで表す. Bag-of-Words とは文書中に出現する単語とその各出現回数のセットで文書をベクトル化するモデルである. この際単語同士の関係性や順番などを考慮しないため, 文章の情報が欠損する代わりに単純で扱いやすいモデルとなっている. 以上の処理によりユーザの書き込みデータを特徴量と顔文字ラベルで表し, これらを教師あり学習に用いる訓練データとする.

3.3.2 教師あり学習による感情推定のモデル

作成した訓練データを用いて教師あり学習による分類器を作成する. 以下では, 本研究で用いる教師あり学習のアルゴリズムである NaiveBayes 分類器の説明を記す [33]. 与えられた特徴量 F に対しクラス C を取る確率は, ベイズの定理を用いて以下の式で表される.

$$p(C|F_1, ..., F_n) = \frac{p(C)p(F_1, ..., F_n|C)}{p(F_1, ..., F_n)} \tag{3.6}$$

分母はクラス変数 C に依存しないため, 分子のみ考慮すれば良い. 単純ベイズ分類器ではここで各特徴量 F_i が他の特徴量 $F_j(j \neq i)$ と独立であるという仮定を用いて,

$$p(F_i|C, F_j) = p(F_i|C) \tag{3.7}$$

が成り立つとする. すると式 3.6 の分子は

$$\begin{aligned} p(C)p(F_1,...,F_n|C) &= p(C)p(F_1|C)p(F_2|C)p(F_1|C)..... \\ &= p(C)\prod_{i=1}^{n} p(F_i|C) \end{aligned} \tag{3.8}$$

と表される. このモデルはクラス事前確率 $p(C)$ と独立確率分布 $p(F_i|C)$ に別れているため扱いやすく, それ故よく用いられる手法である. NaiveBayes 分類器では上記の計算式を利用し各クラスに対し入力データがクラスに属する確率を求め, 最も確率の高いクラスを分類結果として出力する.

本研究の手法において特徴量が単語, クラスは顔文字ラベルとなり, 分類器は入力したテキストが各クラスに属する確率を求める. 特徴量 (単語)F_i がクラス j に出現する確率 $p(F_i|C_j)$ は

$$p(F_i|C_j) = \frac{x_{ij}+1}{\sum_{i=1}^{n}(x_{ij}+1)} \tag{3.9}$$

と書ける. ここで x_ij はクラス j における単語 i の出現回数である. またクラス事前確率は

$$p(C_j) = \frac{t_j}{\sum_{j=1}^{N}(t_j)} \tag{3.10}$$

と書ける. ここで t_j はクラス j に属する訓練データ数, N はクラスの数である. また式 3.9 では全ての単語の出現回数の初期値として 1 を設定するラプラススムージングと呼ばれるスムージングを行っている.$\prod_{i=1}^{n} p(F_i|C_j)$ を計算する際に, クラス j に一度も現れない単語があったとすると $\prod_{i=1}^{n} p(F_i|C_j) = 0$ となり, 他の単語の出現頻度に関わらず $p(C_j|F_1,.....,F_n) = 0$ となる. これはゼロ頻度問題と呼ばれる不具合であり, これを回避する為にスムージングが必要となる.

以上の計算式を用いて, 入力されたテキストがクラス j に属する確率を

$$p(C_j|F_1,.....,F_n) = p(C_j)\prod_{i=1}^{n} p(F_i|C_j) \tag{3.11}$$

と計算する.

分類器の評価は交差検証による精度評価とする. 交差検証とは標本となるデータの内の一部を評価用データとして残りのデータで訓練を行う手法で, 本研究では K 分割交差検証を用いる. まず訓練データを K 個に分割し, (K-1) 個のデータで訓練させ残りの 1 個で評価するという試行を K 回 (分割された K 個のデータ群がそれぞれ一回ずつ評価用データに使われるようにする) 行う. また分類器の評価値として主に Accuracy, Precision, Recall, F-measure の 4 種類が用いられる. Positive クラスと Negative クラスに分類する 2 クラス分類器を仮定すると, Positive クラスについての各指標値は

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.12}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.13}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.14}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3.15}$$

と計算される. 但し, TP(True Positive) は正しく Positive クラスに分類された数, TN(True Negative) は正しく Negative クラスに分類された数, FP(False Positive) は誤って Positive クラスに分類された数, FN(False Negative) は誤って Negative クラスに分類された数を表す. Accuracy は分類の正答率を, Precision はクラスに分類されたデータのうち正しいクラスのデータの割合を, Recall はあるクラスのデータのうち正しく分類されたデータの割合を表す. また F-measure は Precision と Recall の調和平均を表す. 二つの値が等しい場合のみ算術平均を一致し, それ以外の場合は算術平均よりも低い値を取る指標になっており, Precision と Recall のバランスを示す指標として扱われる. 本研究では多クラス分類器には Accuracy を, 2 クラス分類器には Accuracy, Precision, Recall, F-measure を算出し, 評価する.

第 4 章

ソーシャルメディアにおける顔文字分類

4.1 実験の目的

この章はソーシャルメディアにおける日本語テキストの顔文字を対象に, 顔文字の自動分類と感情ラベルに資する情報の抽出を目的とする. まず大規模なユーザの書き込みデータから, 分散表現モデルを用いて顔文字の使われるコンテクストをベクトルで表し, これをクラスタリングし, クラスタの評価関数である DB 値を用いて評価する. 次に, 分類結果及び顔文字と共に使われている単語から顔文字のラベル表現の抽出を行う.

4.2 各データについて

顔文字の分類に用いるデータは, 2012 年 1 月 1 日から 2012 年 3 月 31 日における日本語の Tweet の内 10% を取得したものとし, その中で出現頻度の高い順に抽出した 1000 個とする[*1]. また word2vec に学習させるデータは顔文字の抽出と同じデータとし, 入力するパラメータは 4.1 にまとめた. 次に, K-Means クラスタリングは Python の機械学習ライブラリである

表 4.1. word2vec パラメータ

パラメータ	分散表現の次元	モデル	出力層の近似法	ウィンドウサイズ
値	300	CBoW モデル	階層的ソフトマックス	前後 5 単語

scikit-learn に実装されているものを利用する[*2]. クラスタ数は後述するように 2 から 29 とし, 反復回数は最大 300 回とした. またベクトル距離はユークリッド距離を用いた. 最後に顔文字と共に用いられている単語抽出に用いたデータについて記す. 2012 年 1 月 1 日から 2012 年 12 月 31 日までの Tweet から対象としている顔文字を含む Tweet を一部取得し, その中か

[*1] 抽出に際し, 特殊記号の関係で 2 つの顔文字を除き, 実際に実験に用いている顔文字は 998 個となっている.
[*2] http://scikit-learn.org/stable/index.html

ら各クラスタ最大 5 万 Tweet ずつとなるように顔文字の分類結果に則して Tweet を各顔文字クラスタ毎に分類した.

4.3　結果と考察

図 4.1 は 2 個から 29 個の各クラスタ数毎に 10 回ずつクラスタリングを行い, 各試行毎に DB 値の平均をプロットしたものである. DB 値は値が小さい程クラスタの分割精度が良い事を表す. 全体を通して見ると 5 クラスタに分類した場合が最も指標が良くなっており, 次いで 14 クラスタに分類した際が良い指標値になっている. そこで, 以下の検証では 5 クラスタに分類した場合と 14 クラスタに分類した場合を分類結果とし, 検証することにする.

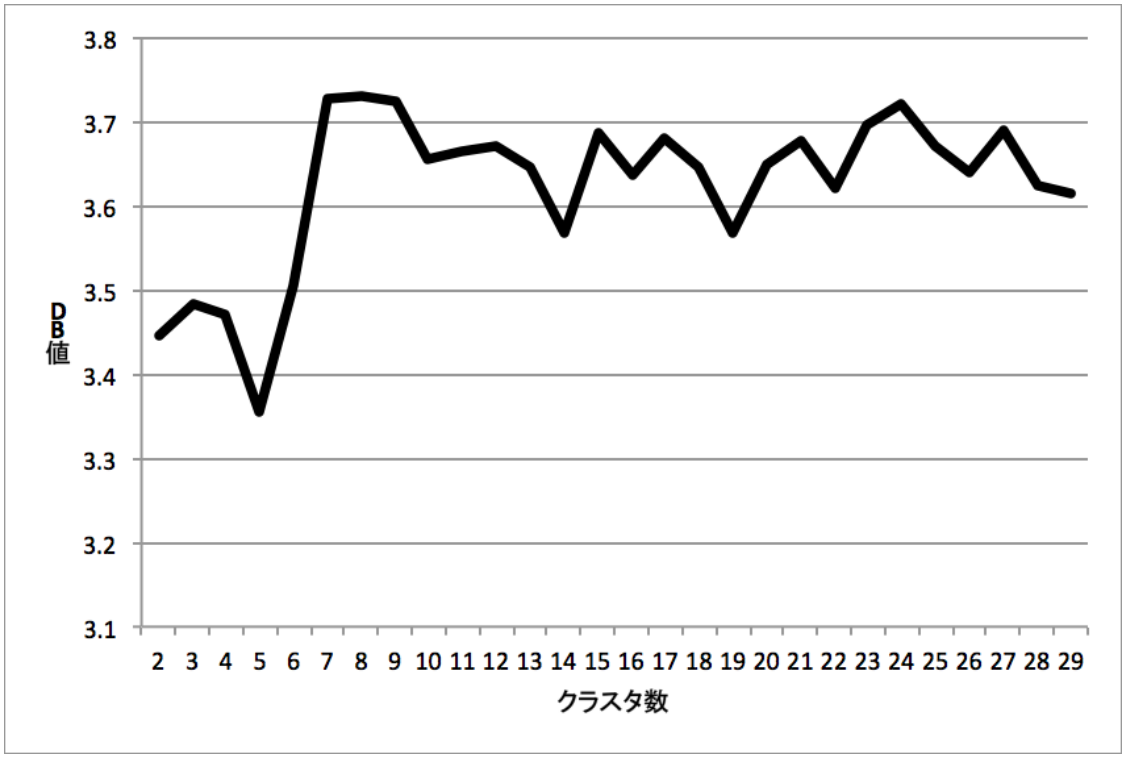

図 4.1. クラスタ数毎の DB 値

4.3.1　5 クラスタ分類について

5 クラスタ分類について, 表 4.2 に出現頻度の高い顔文字をクラスタ毎にまとめた. クラスタ A の顔文字を見ると, 悲しい顔をしている顔文字及び焦った表情を表す顔文字が多く, 「軽い Negative」を表す顔文字のクラスタと思われる. クラスタ B には泣いている顔文字が多く, 「強い Negative」を表す顔文字のクラスタと思われる. クラスタ C には笑顔で喜んでいる顔文字が多く, 「強い Positive」を表す顔文字のクラスタと思われる. クラスタ D にも笑顔の顔文字があるが, 手を挙げる等あいさつをしている顔文字が多く, 「軽い Positive」を表す顔文字の

クラスタと思われる. また, クラスタ E の顔文字については共通する傾向があるかどうかは分からないように思われる.

表 4.2. 5 クラスタに分類した顔文字

A	B	C	D	E
(´･_･`)	(>_<)	＼(^o^)／	(^^)	(`・ω・´)
(´Д`)	(T_T)	(((o(*°▽°*)o)))	(*^^*)	(・∀・)
(´・ω・`)	(;_;)	(*´ω`*)	(´ ▽ `)ﾉ	(`・ω・´)
(´･ω･`)	ヽ(;▽;)ノ	(*´▽`*)	(^−^)	(・ω・)
(￣▽￣)	(´･ω･`)	(^q^)	(^o^)	(°∀°)
(^_^;)	(´;ω;`)	(*´∀`*)	(^−^)/	(≧▽≦)o
(´Д`)	(´;ω;`)	¥(//▽//)¥	(^O^)	☆*:.。
(´・ω・`)	(´;ω;`)	＼(^O^)／	♪(´▽`)	。.:*☆
((((;°Д°))))))))	(T−T)	(^ω^)	o(^▽^)o	(・ω・)
(*_*)	__|￣|○	(・∀・)	(^_^)	＼('ω')／
(。−_−。)	(;__;)	(´ω`)	(*^o^*)	(≧▽≦)
(￣—￣)	(;o;)	(´∀`)	(*^_^*)	(`・ω・´)
(;´Д`)	(´;ω;`)	(´ ▽ `)	(^O^)／	(。・ω・。)
(−_−;)	。・°・(ノД`)・°・。	...。.:*・'(*°▽°*)'・*:.。	(o^^o)	(´−`).。
(^^;;	(/ _ ;)	(*´ω`*)	♪(´ε`)	ヽ(^o^)
(;´д`)	／(^o^)＼	(￣▽￣)	(*`ω´)	///」
(^^;	。°(°´Д`°)°。	(*´▽`*)	＼(^^)／	%) (
(−_−)	(; ;)	(≧▽≦)	m(_ _)m	Λ('Θ')Λ
(￣▽￣;)	(T^T)	(*´∀`)	(^^)/	(^ω^)
(￣∀￣)	; ∀ ;)	(*´д`*)	↘(@￣—￣@)ノ	<●><●>
(^−^;	(ノД`)	(*´Д`*)	(^ ^)	————(°∀°)——
(^_^;)	(;ω;)	☆〜(ゝ。∂)	o(^−^)o	((_ _))..zzzZZ
(^^;)	(T__T)	(´▽`)	(#^.^#)	(((^−^)))
(−−;)	(;_;)	(*≧∀≦*)	m(_)m	(`・ω・´)
(´д`)	(/_;)	)^o^(	(^o^)/	(ﾉ◡ﾉ)♡
(￣◇￣;)	(´° .....ω° .....`)	(´∀`*)	(^ з ^)−☆	(､ン､)__
(;´Д`)	(><)	(●´ω`●)	ヽ(･∀･)ノ	=(^o^)=

次に感情ラベルに資する情報を得るため, 各クラスタの顔文字と共に使用されている特徴語を見ていく. 表 4.3, 表 4.4, 表 4.5 に出現回数, TFIDF 値, 指標値 $R_j(i)$ の各抽出法に基づいてスコアの高い順に単語を 50 個ずつ載せた. まず出現回数を用いて抽出した特徴語 (表 4.3) を見ると, 「て」や「や」等, ラベル表現の抽出の観点から考えて目的に合わない単語しか出ていない. 次に TFIDF を用いて抽出した特徴語 (表 4.4) を見ると, 各クラスタで同じ感情の傾向を持つ単語がある事が分かる. 「Negative」と判断したクラスタ A と B には「寂しかっ」, 「キツイ」, 「ミス」, 「ぇぇ」等の Negative を表す単語が出てきており, 「Positive」と判断したクラスタ C と D には, 「楽しも」, 「おもしろかっ」, 「成功」, 「余裕 」等の Positive を表す単語が出てきている. しかし, クラスタ E は特徴語に同じ感情の傾向がないと思われる. 最後に指標値 $R_j(i)$ を用いて抽出した特徴語 (表 4.5) を見ると, 短く端的に感情を表す単

語が多く, また感情と関係のない単語が少ないことが分かる. クラスタ B の「悲しい」, 「泣く」, 「ううう」, 「痛い」等はクラスタ A の「きつい」, 「悪い」, 「苦手」, 「困る」等より強い Negative を表す単語のように見える. またクラスタ C は「最高」, 「おいしい」, 「楽しみ」等の強い Positive を表す単語があり, クラスタ D は「こんにちは」, 「はじめまして」, 「フォロー」, 「どうぞ」等のあいさつを表す単語がある. しかし, クラスタ E は特徴語に同じ感情の傾向がないと思われる.

TFIDF 及び指標値 $R_j(i)$ により抽出した特徴語は, 顔文字の一覧から捉えられる特徴と一致しており, 特に指標値 $R_j(i)$ を用いた場合は一般的に感情語と考えられる単語が多く出てくるため, 共起語から各クラスタの感情ラベルに資すると思われる.

表 4.3. 5 クラスタ分類の各クラスタの特徴語 (単純回数)

A	たてのにはねんないよだなですがかーでもしとから けどそうてるます笑ってお私やをたらいいうとかいわことそれえまし さ人さんねーたいなっれじゃあある
B	てたのにんねよないですーはだがでなもしかとから ますそうけどお私うありがとうてるたいいいをやたらって笑いましえさんなっ さわあよーとかれああそれちゃんねー
C	てたのーにはよねですでだしもんながかおますと ないからいい笑そうてるうありがとうさんいよーましちゃんやございけど私たらをたい わさってきあねーっｗとかえ
D	てたのーはによですねでしだますもんなおがかと からないありがとう笑いいございよーましてるそううさんいをたらけど私やさあり ちゃんってきたいわあ今日ねーあるありがと
E	てたのにはーよですねでだんなしもがかないとます からおてるう笑いいそうをいさんけど私ってたらやさましよーわｏ ちゃんたいありがとうえあｗことっきござい

4.3.2 14 クラスタ分類について

14 クラスタ分類について, 表 4.6 に 5 クラスタと同様の方法で顔文字をまとめた. クラスタ 0 は手を挙げてあいさつをしている顔文字が多く, クラスタ 1 は震えて怯えている顔文字が多い. クラスタ 4 は眉毛をあげて顔に力を入れた顔文字が多く, クラスタ 5 は目を細めて笑顔の顔文字が多い. クラスタ 6 は睡眠を表す顔文字や誤っている顔文字が多く, クラスタ 7 は強く喜んでいる顔文字が多い. クラスタ 8 は泣いている顔文字が多く, クラスタ 9 は悲しんでいる顔文字が多い. クラスタ 10 はクラスタ 0 同様に手を挙げて呼びかけておりかつ表情の柔らかい顔文字が多く, クラスタ 11 は焦っている顔文字が多い. クラスタ 13 は泣いている顔文字が

表 4.4. 5 クラスタ分類の各クラスタの特徴語 (TFIDF)

A	げろ 揺れ ょぶか どど 引退 暑かっ なくなる 何故か イマイチ 無く うける のど 多かっ 恐い 動 辛かっ 返せ 削除 寂しかっ レッスン 切なく 詳細 出来 る ノリ 映像 聴き 聴い 余裕 死ん もらえ もらう もらい 日本人 落ち着い もらっ ぇぇ 院 午後 だ 早速 ミス すでに キツイ ぇっ ちまっ 簡単 まぢか
B	さびし 惜しかっ ズル 揺れ 切な キャンセル ぁあ ツライ 寂しかっ 無く ほしく 有難う 返せ 限っ 届か 落選 全滅 去っ ｗｗｗｗｗｗｗｗ 暑かっ 削除 恐い 想い 出来 る もらっ 映像 聴き 聴い 余裕 死ん もらえ もらう もらい 日本人 ばあちゃん 落ち着い インフル ぇぇ 院 午後 だ ミス すでに 無くなっ キツイ
C	忍び込む れいれい 楽しも すさ リフォロー うける 有難う ランチ めで 堪能 何故か 宿 出来 る もらっ 聴き 聴い 余裕 死ん もらお もらえ もらう もらい 人生 落ち着い あおい 午後 だ 早速 昼寝 すでに おもしろかっ くださ 簡単 おもしろ 個 あやか 5 枚 無 全力 みなさん 布団 成功 引き 検索 歩い 仲良し 運
D	リフォロー ヨロシク 有難う ランチ 呼べ うける どーぞ 鳥 たのしん 落ち着く 出来 る もらっ 聴く 聴き 聴い 余裕 死ん もらえ もらう もらい ばあちゃん 落ち着い 科 ぇい 午後 だ 早速 すでに おもしろかっ くださ 簡単 まぢか 個 5 枚 全力 みなさん 布団 引き 検索 歩い 仲良し 運 いっしょ 乗り 也 死ぬ
E	ー アア 卍 縮む 有難う うける リフォロー 呼べ ｗｗｗｗｗｗｗｗ 桃 判断 任せろ 諦める 管理 ずっ どど 無く ぶる 詳細 出来 る もらっ 聴き あがる 聴い 余裕 死ん 晒し もらお もらえ もらう もらい 魔法 ばあちゃん 落ち着い 科 ぇい 午後 だ 早速 昼寝 すでに くださ 簡単 おもしろ 個 5

多いが, クラスタ 8 の方が涙が強調されている. またクラスタ 12 はネットスラングでよく用いられる顔文字が集まっており, クラスタ 2 と 3 は詳細な傾向は判断できないと思われる.

次に感情ラベルに資する情報を得るため, 各クラスタの顔文字と共に使用されている特徴語を見ていく. 表 4.7, 表 4.8 には TFIDF による単語抽出の結果を, 表 4.9, 表 4.10 には指標値 $R_j(i)$ による単語抽出の結果をまとめた. まず TFIDF を用いて抽出した特徴語 (4.7, 表 4.8) を見ると, 地名や人の名前等の固有名詞や, 「ぃいいいいいいいいいいいいいいい」「オォオオ

表 4.5. 5 クラスタ分類の各クラスタの特徴語 ($R_j(i)$)

A	f うーん 暑い 微妙 怖い 最悪 なんだか 悪く 少ない けっこう 難しい 不安 大変 困る 厳しい まったく 苦手 変 なぜ 嫌 こわい おかしい ダメ きつい 実際 たしかに 長い 迷惑 会社 わから 謎 強い 内容 しまう なかなか 悪い 差 ビックリ しょうが あまり わかん バカ 急 登録 生活 そんなに 余計 なんで かかる 病院
B	うぅ 悲しい t 泣く tot 泣き ショック 泣い 涙 ごめんなさい 寂しい ううう つらい つら 痛い たかっ 泣 厳しい 申し訳 ごめん 残念 すいません 辛い 遠い ああ うわ ひどい 行け 感動 遅れ すみません 痛 よぅ ひい 遅く 不安 おおお きつい 心配 あかん 緊張 しまっ 消え 余計 すま こわい 体調 のに 忙しい 寒い
C	照れ おいしい たのしみ わーい 美味しかっ 嬉しかっ しま 最高 ほっ 可愛かっ 癒さ おいし おめでとう 呼び 美味しい かわいい 楽し ぜひ 歓迎 ゆい ふふふ いこ 楽しん 可愛い 楽しみ わい うれしい め 誕生 大好き 面白かっ 翔 素敵 あざ へ 可愛 アイコン 楽しかっ 綺麗 もちろん 艸 あげる つつ 仲間 かっこいい 萌え 美味し 懐かしい もも 嬉しい
D	q こんにちは らっしゃい いたし はじめまして おや てらっ フォロバ 了解 よろしく 宜しく フォロー てら はーい どうぞ こちら おはよう お願い 仲良く 楽しかっ おか こんばんは いえいえ お気 こそ 是非 歓迎 ござい 遊ぼ いただき 頂き おいで タメ あけ お疲れ様 お疲れさま 楽しん 呼ぶ 呼ん 今度 おやすみ あり おかえり もらい しゃ ただいま 面白かっ こん 下さい つかれ
E	d 乙 o o メモ 三 ノ おやすみなさい 眠い おやすみ うむ wwwwww v がん m パン ぽん がんばり 基本 っと こら ぜ ド 絡み がんばれ 全力 9 なるほど きゃー 寝る エロ 萌え 会お ぼっ 報告 せる wwwwww おう なさい しっかり みよ！用 とり へー ぞ ろ ほう

オォオオォオオオアアァァアアァァア」等の口語ベースで単体で意味を成さないくだけた表現が多く, ラベル表現の抽出の観点から考えて目的に合わない単語が多い. 次に指標値 $R_j(i)$ を用いて抽出した特徴語 (表 4.9, 表 4.10) を見ると, 口語ベースのくだけた表現もあるが, TFIDF の場合より端的に感情を表す単語が多い. クラスタ 0, 10 は「こんばんは」, 「了解」, 「どういたしまして」, 「サンキュー」等のあいさつを表す単語があり, さらにクラスタ 10 では「ワクワク」, 「楽しん」等の Positive を表す単語もあり, それぞれ顔文字の一覧から

表 4.6. 14 クラスタに分類した顔文字

0	1	2	3	4
(￣^￣)ゞ	((((;゜Д゜))))))))	(￣▽￣)	(≧▽≦)o	(`・ω・´)
(^o^)／	(◎_◎;)	(^q^)	☆*:.。	(・∀・)
(`_´)ゞ	((((;゜Д゜)))	(・∀・)	。.:*☆	(`・ω・´)
(^^ゞ	(°口°)	(*`ω´)	＼(ˈωˈ)／	(・ω・)
(^_^)/	Σ(・□・;)	(。-_-。)	(≧▽≦)	(°∀°)
(*・ω・)ﾉ	Σ(￣。￣ﾉ)ﾉ	(￣ー￣)	ヽ(^o^)	(^p^)
(*°▽°)ﾉ	(°　□°　;)	(^▽^)	///」	(`・ω・´)
(・∀・)ﾉ	(;*°ˈω°*)	(・ω・)	%)　(	(。・ω・。)
(^∀^)ノ	(°▵°)	(￣∀￣)	(ˈ◡ˈ)♡	(^ω^)
(・∀・)ノ	(￣□￣;)	☆～(ゝ。∂)	(ノ∀`)	<●><●>
(^_^)ノ	((((;゜Д゜))))	(^∇^)	＼(^o^)／	(`・ω・´)
(^−^*)/	(°口°;	)^o^(	≫(。゜∀゜)ﾉ･ﾟﾟ･。	(,ン,)__
(^−^ゞ	(((・・;)	ψ(`▽´)ψ	(o^▽^o)ﾉ	=(^o^)=
(`・ω・´)ゞ	(°　Д°　)	⊂((・x・))⊃	_T)＼(^−^)	(´^ω^`)
(。・ω・)ﾉ	(;・∀・)	(￣▽￣)	(o^−')	(・∀・)

5	6	7	8	9
(*´ω`*)	m(_ _)m	(((o(*°▽°*)o)))	ヽ(;▽;)ノ	(´・ω・`)
(*´▽`*)	m(_)m	¥(//▽//)¥	(´;ω;`)	(´・ω・`)
(*´∀`*)	(−_−)	*･゜゜･*:.。..:*･'(*°▽	(´;ω;`)	(´Д`)
＼(^O^)／	(−.−)	(*´д`*)	(´;ω;`)	(´・ω・`)
(^ω^)	(´−`).。	(*´Д`*)	(´;ω;`)	(´・ω・`)
(´ω`)	((_ _))..zzzZZ	(//▽//)	__|￣|○	／(^o^)＼
(´∀`)	m(._.)m	♪———O(≧▽≦	(;__;)	(´д`)
(*´ω`*)	<(_ _)>	＼(^o^)／♡	(´;ω;`)	(´・ω・)
(*´▽`*)	(_)	(*°▽°*)	。･゜･(ノД`)･゜･。	(´・ω・`)
(*´∀`)	(´д⊂)	(*/ω＼*)	。°(°´Д`°)°。	(´−ω−`)
(´▽`)	(つ∀−)	(〃▽〃)	(;ω;)	(´`)
(´∀`*)	(_ _)	(〃▽〃)	(´°　.....ω°　.....`)	(´Д`)
(●´ω`●)	(σ ω−)。o	(*´Д`)	。･゜°　･(>_<)･°	(･ω・`)
(^ω^)	(=_=)	(≧▽≦*)	(; ω ;)	('A`)
(´ω`)	_(._.)_	(*´д`*)	(;ﾟ∥Д∥ﾟ)	(((^−^)))

10	11	12	13	
＼(^o^)／	(´Д`)		(>_<)	
(^^)	(^_^;)		(´･_･`)	
(*^^*)	(*_*)		(T_T)	
(´ ▽ `)ﾉ	(;´Д`)	————(°∀°)————	(;_;)	
(^−^)	(−_−;)	_φ(･_･	(T−T)	
(^o^)	(^^;;	■■■■■■■ ★	(;o;)	
(^−^)/	(;´д`)	]／“	(/ _ ;)	
(^O^)	(^^;	■■■ ★	(; ;)	
♪(´▽`)	(￣▽￣;)	—(°∀°)—	(T^T)	
o(^▽^)o	(^−^;	————(°∀°)————	(´Δ`)	
(^_^)	(^_~;)	φ(. .)	; ∀ ;)	
(*^o^*)	(^^;)	φ(..)	(ノД`)	
(*^_^*)	(−−;)	————————(°∀°)——	(..)	
(^O^)／	(￣◇￣;)	φ(..)	(T__T)	
(o^^o)	(;´Д`)	——(°∀°)——	(:_;)	
		——(°∀°)——		
		[B!]		
		φ(°−°=)		

捉えられる特徴と一致している. クラスタ 1 では「恐ろしい」, 「ビックリ」等, クラスタ 9 では「寂しい」,「悲しい」等, クラスタ 11 では「めんどい」, 「焦る」等の Negative な単語が多く, それぞれ顔文字の一覧から捉えられる特徴と一致している. クラスタ 8, 13 では「泣ける」, 「号泣」, 「ひどい」, 「涙」等の強い Negative を表す単語が多く, それぞれ顔文字の一覧から捉えられる特徴と一致している. クラスタ 5, 7 では「可愛い」, 「幸せ」, 「うれし」, 「ワクワク」等の強い Positive を表す単語が多く, それぞれ顔文字の一覧から捉えられる特徴と一致している. クラスタ 6 は「すみません」, 「ごめんなさい」, 「お手数」,「申し訳」等の謝罪を表す単語が多く, またクラスタ 2, 3, 4, 12 は感情を表す単語が少ないと思われる.

4.3.3 5 クラスタ分類と 14 クラスタ分類の比較

表 4.11 は 14 クラスタの各顔文字が 5 クラスタ分類ではどのクラスタに分類されているか集計したものである.

表 4.11 からクラスタ A は 14 クラスタ分類のクラスタ 1,9,11 の顔文字が, クラスタ B は 14 クラスタ分類のクラスタ 8 及び 13 の顔文字が, クラスタ C は 14 クラスタ分類のクラスタ 5,7 の顔文字が, クラスタ D は 14 クラスタ分類のクラスタ 0,10 の顔文字が, クラスタ E は 14 クラスタ分類のクラスタ 3,4,12 の顔文字が対応している事が分かる. またクラスタ 2 とクラスタ 6 は一つのクラスタには対応していない事が分かる. 前節及び前々節の結果を踏まえると, クラスタ A, B, C, D に対応するクラスタはそれぞれ似た感情語を持つ 14 クラスタ分類のクラスタが対応しており, 感情語を抽出できなかったクラスタ 3, 4, 12 がクラスタ E に対応している事が分かる. ここから, クラスタ E は Positive でも Negative でもない顔文字が集まっており, Neutral を表すと推察できる. また, クラスタ 6 に属する `m(__)m` や `m(_ _)m` の顔文字は Negative を表すとも解釈できるが, 5 クラスタ分類では, クラスタ 0, 10 の顔文字と共にあいさつを表すクラスタ D に属している. これは江村らのいう, 感情よりも動作を表すとするコミュニケーションタイプに位置する顔文字クラスタだと考えられ, 先行研究の結果と親和性のある結果となっていると推察される [14].

表 4.7. 14 クラスタ分類の各クラスタの特徴語 (TFIDF) 前半

0	グセ スミマセーン アレッス スミマセーンウチ ゃんおざ マジギレ ぷさんおざ olick ぉはゅん ゃんおはよろ ヽ マチャキ waka 敬礼 ぴさんおはよろ 獅子王 虎屋 三河屋 ショウ ぃちゃんおはよう スクラム 歩美 ニョッキ ぇっさ ミレイ 出戻り 千晴 瑠璃 ゃんおつかれさま ミィ 三平 ハーイ ゆきみ やまや ふじょ 取り扱っ ざぁす てらさ ボウ 小梅 ロデオ ラジャー マサキ ょかい 健闘 らっし しさん
1	おそろし ガーン 費やし 驚愕 危なかっ ぃぃぃぃ 行列 マズ 制度 爆弾 ぇぇぇぇ なんてこった 光っ コワイ 自覚 いつのまに トラウマ パク ぁぁぁぁぁ 殺す 受ける 罠 バク たいへん 追いかけ わるい ゴキブリ 限っ ずいぶん 曇り おなじく 物理 遠足 なんぞ 凄かっ 聴き 重症 うっかり 切実 メッチャ ス 仲 本名 あんた オーラ すっぴん
2	ソンギュ 黙 オト 麻衣 ペロペロ せこい 苛 アゴ 住む ひろこ ｉ ゆかた 気持ちいい えぐい えれ やれる おもしろかっ 起こさ 世話 返せ 協力 ウイルス 惚気 目指す ’マオ ねッ 個別 病み くるん 受ける わくわく トコ 睡魔 信用 ちょいちょい 連続 なつき さりげ メッチャ 薄い おさん ス 本名 虫 相性 仮 jk
3	ぃちょん feat ウオオオアアア wwwwwwwwwwwwwwwww シャシャ ーキャラカブ ウオ 鞘 もしもし crazy オオ オォオオォオオォオオオアアァアアァアア ゃんた 成嶋 アア wo ざいますみやと ｗｗｗｗｗｗｗｗｗｗ ナノ 父 お決まり キャッ シフ ぱちぱち ぁいちん テニ 那 縮む 止まれ けんさん ＜ 天馬 ヨシ 美少女 ずん わいわい 楽しも ＞ ｉ ｑ ミッション はろ ぁいてぃ 弥 正常 蘭
4	パシッ ｑ 江田島 平八 くわわっ ぺいはうれしそうにばしゃにかけこんだ 三河屋 ヒナギク eight × eighter 汚名 エド 平が 駆け込ん ロイ 馬車 三平 喜多 加わっ プ 名乗る gogo いかに 乗り切る 障 精進 ● 海賊 ジョジョ 素晴らし ねぎ ff ー 帝 ぼう むかつく ｑ 煙草 どんと 水戸 やれる リク だるま してやっ イヤイヤ 誇り 後期
5	sacbys weaver しんたろう ぇようにないってら ジョナ ぁりがとぅ 喜多 ゥ ィィ あいさ ナツ 勲 和む サクラ なァ 萌子 むふふ あたる ゃんおはあり めで 隆二 申し ＞ wwwwwwwwwww 一安心 此方 初代 ィ エリック オムライス トキ 楽しま ふろ みい ｏｋ 森田 me ぉろ ねる 気長 たげ 滝 たまり 協力 呼ぼ いらっしゃい 眺め 番目 悟
6	恐れ入り 日笠 申し上げ 何卒 志賀 本年 当方 ｂａｂｙ emiko ぁりがとうございます 豊崎 レイン スマン 冥福 ミィ 誠に 夜分 原発 オヤスミ やゆ お陰様 遅ればせながら 取り置き 移行 遣い 眠かっ ありがたき グァンス 氷室 節 トモ オ かしこまり 先ほど くれぐれも お越し 返 昨年 お送り 是非とも ガンバっ 自愛 ゴメンナサイ ご覧 スパム お詫び

表 4.8. 14 クラスタ分類の各クラスタの特徴語 (TFIDF) 後半

7	ぴば sunao 伊月 すう ろば おめおめ はっぴ かっこよさ するする ぉぉお 臣 介 佐々木 じゃい 神田 真夏 北山 増田 トキ 沖田 ピュア ぜひとも チャミ 格好いい そっくり フェチ おもしろ おうおう パチ カッコイイ 爽やか キュンキュン 光栄 揃い 川越 かっこいー 愛しい 将 絡も 泊り ウヨン たまり 会し ぁぁぁぁぁ トゥギ 協力 げろ 手越 あわわ
8	ぇは ・ さびし 凹 三平 キュヒョン sunao ぁぁん サヨナラ 慎吾 ゞ ぐす ょぶか うわん ゲート 健二郎 ええい ぁぁぁあ 見れん ぁぁぁん ひじ 在庫 誘え 探さ 冷め 中居 行けん ねえさん 代行 逢え 断念 下さっ エラー 座席 咳 ぃぃぃ 追わ 暁 愛しい 返せ 病む 亜 wwwwwwwww 少な 発送 なさっ ムカ 何だか 確率
9	菜美 三平 つかまっ 誤魔化せ ー しょんぼり こる 喋ら しょぼ ＞ エリック もやもや 通行 寝かせ 続か 壊す 曇っ 喋れ ねむ 捨てる 悲しく url 童顔 一般人 無くなっ へン 待た 華 wwwwwwwww 無言 言 キャベツ ボカロ 女子高 難しく 之 おとなしく 略し 曜日 武 宮地 悪化 赤司 のせ 説 何だか 鬱 すんません
10	ゃんおはよかい ぇんりるさんおはありの ぃぃぃぃぃぃぃぃぃぃぃぃぃぃぃぃ ぃぃぃぃぃぃぃぃぃぃぃぃ ジミー ぁりがとうございます ファイティング テガン ぁりがとう 克 ウノ おおおおおおお おおおおおおおお tk ローサ thankyou 島根 ガンバっ テル 遊佐 護 絡も レオ ぉございます 候補 スノボ 解消 ジュン ぉです ともさ あそび 苺 りんも バド くるん たのしん うれし ホワイト 黒い 津 ゆみこ じゅんち だて 瀬戸 ジュンス 行く行く 日向 陸上 くださ
11	ピィ shin 災難 危なかっ 阪急 ミル 習慣 バタ ゆたか マズイ 減り 様子見 悲惨 続か ノロ 憂鬱 警報 個別 しんどかっ 冷め 長男 愛媛 えぐい 古典 いんす kg 止ん 釣れ タル ビビり ダラダラ オイル みわ 落ち着か 病む ひどく えら 食え 腹立つ 暖房 経営 暗く 気温 当分 吐き ロケ 期限 帰ら
12	ー ー リプ フムフム ぅほぅ ぉほぉ カキ ふじこ 哲学 圧 vv 解除 wqt 新作 wwwwww 眼鏡 外 もらい 範囲 恥ずかしい 血 セブン 企画 ぶり 変わっ うまく きち 恋 彼 シーン トイレ とにかく とても ブログ 秋 外れ ポイント v きゃー すげ
13	火事 じんましん joy ジャイ 日誌 痒い 瑞 ノロ 切なく 無くなっ 治り 古典 ツライ 細 咳 苦労 ガッカリ 落選 ひど シングル 事務 図書館 寄っ 現象 ハズレ さむい なんとも 萎え 気温 すんません 大人しく 混む 恋しい たいへん あいた 点滴 smt 落ち着か 返せ 待た 腫れ 名義 未成年 ながい コム 腹痛 eighter 下がる

表 4.9. 14 クラスタ分類の各クラスタの特徴語 ($R_j(i)$) 前半

0	一足 配達 ざ こんばんは 寝坊 よろ bot てらっ らっしゃい 隊長 いやぁ 了解 三平 ゃい おはよう ス やみ だて いて お気 三河屋 終了 ジェジュン りょうかい こんち 多く ゃんおはよ こんにちは けい しゃ らん 昨夜 おさん ボク 桜 なつ くす 完了 おや はお ご苦労 美 どういたしまして 辺 まま まる 本日 す ほか くり
1	恐ろしい どど 怖 あわわ 恐い こわ 恐怖 ぎゃ 怖い 何故 ぇぇ えっ びびっ えーっ 衝撃 こわい ビックリ ぇぇぇ まさか ホラー そそう もしかして びっくり そそ べし なぜ ぁぁぁぁ ぎゃあ え 危ない 雷 うそ 怖く 初耳 危険 すげ すかっ はっ そんなに e すご 工 キロ 食べ物 どういう やば ええ 焦っ もしや やめろ
2	すてき あんた たる おごっ お前 うる ぺろぺろ へっ おもろい ふっ 太る ひひ おまえ 笑ま し笑 はげ 青春 あげよ 暇人 勝ち 爆笑 内緒 狙っ 黙っ たれ こら 取る にやにや バカ ウケる ふふふ わら 自慢 笑も 教科書 ざま 怒っ 例 オッケー はあ あいつ ばか なめ 許さ チャラ くせ 美味い がり ハマっ たのし
3	wwwwwwwwww q ぁいちん t o o 剛 電 j . ー 豆腐 ぺろ みや 卵 える さっち wwwwwww わーい ぁいと おめでとう wwwwwwww くみ ファイト どうも ステキ ぁい ぇい どい ゃんおはよ ペ 晒し ひかる よし 全力 てぇ ♪ 誕生 めし 注文 濡れ ひで 垢 プレイ 減っ ッ wwwww 全裸
4	3 焼きそば 溺愛 牛乳 d ゆりえ o コーヒー ッス 乙 パン 三河屋 ワシ ス wwww もちろん 三 壁 クラスタ 赤司 k wwwwww wwwwwwww wwwww 塾 忘れ 描き 待機 全裸 wwwwwww わろ お呼び 天才 ぶん 尻 オレンジ うむ うさ フラグ 厨 描い ホモ 攻め 描く ござる 目指し おもう 主 うめ
5	sacbys 歓迎 気軽 呼び 癒さ タメ 可愛かっ イラスト リフォ かわゆい まったり 総 御 お礼 うふふ ぺろぺろ かわいかっ おおっ 呼ば こり 幸せ ほっ おいしかっ 可愛い 呼ぶ フォロバ 絡ん 梅 嬉しかっ 萌え 改めて いお 美味い なで 華 なっちゃん こんにちは 仲良く 魅力 キレイ 似合い 気に入っ ふふふ 届き 揚げ 便利 おいし 絵 素敵
6	申し上げ 致し お手数 来店 おかけ 騒がせ おやすみなさい いたし 宜しく お待ち お世話 失礼 今晩 すんません 眠い お願い すみません 御座い 今後 すいません ご苦労 おり 本日 ゴメン 有難う 申し訳 ごめんなさい 助かり 検討 譲っ まして 頂け どうぞ 頂き 丁寧 おやすみ いただき 明け 頂い 願い 突然 有り難う 初め 皆さん 人数 眠く 感謝 いただい 掛け わざわざ

表 4.10. 14 クラスタ分類の各クラスタの特徴語 ($R_j(i)$) 後半

7	照れる ぅぅ きゃー うれし わくわく たのしみ こま きゃ 惚れ かっこ かっこよかっ 淳 きゃあ かっこよ コナン 照れ かっこいい 会お ぶら ゃあ たまら かわい 眼鏡 かわゆい 双子 会える よっしゃ 語ろ 大好き カッコ かっこよく 似合い 嬉し やば ろう ステ ぁいてぃん 可愛かっ やばかっ わーい ワクワク 可愛 気軽 むっちゃ やま 可愛い 最高 すてき すご にやにや
8	号泣 泣ける うぅ 泣け 泣き うう 涙 泣く ー ちゃあ 泣い 寂し 優し 感動 ｉ 悲しい さみしい つらい pq おおお ふお ああ ぁぁぁぁ つら 辛 切ない 倍率 ぎゃあ 寂しい 買え 酷い ほしかっ たかっ よお ひどい 見逃し 泣 かわいそう 譲っ ぇぇぇ うわ 耐え ショック 外れ 交通 心臓 ぉぉ ぇぇ ーーーーー 会い
9	枕 三平 工 イヤ 売り切れ そもそも 似合わ 駄目 タン 合わ 年生 寂しく 通っ どし のう うーん 禁止 ry 少ない 寂しい なにか あんまり 分かん かそ うむ 鬼畜 悲しい 出会い 難し 食欲 まとも ゼミ せめて 欲しかっ 不思議 微妙 悩む 難しい 計算 登場 つらい 離れ めんどくさい 垢 面倒 にくい 残念 否定 放置 六
10	リフォ いーえ 香 サンキュー 似合っ フォロ 到着 ぁいてぃん バスケ pq どういたしまして ワクワク 美味しかっ 御座い ウケる ミン 記念 たのしかっ 語ろ うまかっ 遊ぼ おいしかっ あんに とう おかえりなさい ぁいと 呑み フォロバ ちゃお 頑張れ 楽しめ 楽しん めい 館 みなみ 靴 ぐち ヨロシク 花 リフォロー りえ ばり おかえり じゅん 買い物 おもろい ぇい ゃあ はるか あゆ
11	f キツイ 暑 キロ 混ん ギリギリ 飲ま 寒かっ めんどい 焦る そそう 向こう 始まる アメリカ 最悪 ホンマに 痛 地方 キツ 焦っ 何故か 難しい 片付け 限界 分かん 比べ 我が家 ありゃ 量 虫 悩み かかる 症状 不明 しんどい 模試 ムカ 間に合わ 娘 疲れる だるい 辺り 九州 セリフ ビックリ 体力 暑い お客 それにしても 先週
12	メモ ふむふむ キタ っと なるほど ほほ 程 参考 ほう qt 宇宙 チェック つまり へぇ ド 詳しい 追加 重要 理想 厨 モテ 好み 初耳 票 ツンデレ ついに 変態 チーズ ・ ホモ 勉強 おく がり 塩 イケ ｓ ｍ 把握 北 期 おこ 代表 魚 ヲタ 確定 エロ 検討 全国
13	tot 頭痛 怖かっ 寒かっ ショック 悔しい いそ ずるい 切ない うぅ さみしい 号泣 泣き たかっ 寂しい 病ん 揺れ 送れ 悲しい 戻り 金欠 辛い かわいそう 残念 つらい 炎 泣 泣い 合わ 冷たい 来れ 会わ なくなる 地方 泣ける ｕ ごめん ぅん 行け どし 飲め 寂し 地震 乗れ ょぶ ほしかっ ムリ かかり 寂しく 遅れ

表 4.11. 5 クラスタ分類と 14 クラスタ分類における顔文字の所属の比較

クラスタ	A	B	C	D	E
クラスタ 0	0	0	0	49	6
クラスタ 1	43	0	0	0	6
クラスタ 2	37	0	18	13	10
クラスタ 3	0	1	3	8	55
クラスタ 4	0	0	2	0	105
クラスタ 5	0	0	104	1	3
クラスタ 6	6	0	0	6	13
クラスタ 7	0	4	63	0	1
クラスタ 8	1	60	0	0	0
クラスタ 9	50	8	0	0	14
クラスタ 10	0	0	27	128	0
クラスタ 11	80	0	0	0	0
クラスタ 12	0	0	0	0	15
クラスタ 13	13	45	0	0	0
合計	230	118	217	205	228

第 5 章

機械学習による感情推定

5.1　実験の目的

この章では前章で得られた顔文字分類を用いて, 顔文字ラベルを利用した教師あり学習による感情推定を行う. まず感情推定を行う多クラス分類器と感情極性推定を行う分類器を作成した. 感情推定は顔文字クラスタを全て区別し, クラスタを推定するタスクを行う. 同時に各クラスタ毎の推定の難易度を調べるため, 各クラスタ毎に, そのクラスタに属するか属さないかを推定するタスクを行う (クラスタ数と同数の分類器を作成する). 感情極性推定は似た感情の傾向を持つクラスタ同士を統合する事で, Positive, Negative の 2 クラス分類, 及び Neutral を含めた 3 クラス分類を行う. また精度を向上させるために, 分類器に入力する特徴量を変えて精度を比較する実験を行った.

5.2　各データについて

顔文字とその分類は, 前章で行った 1000 個の顔文字の 5 クラスタ分類及び 14 クラスタ分類の結果を用いる. 訓練データには, 2012 年 1 月 1 日から 2012 年 12 月 31 日に取得した, 対象としている顔文字を含むリプライの Tweet のうちの一部を用いる. 序章で述べたように, ソーシャルメディアにおける顔文字はユーザ同士のコミュニケーションにおいて感情や心境を伝える補助手段であり, Twitter においてユーザ間のインタラクションの手段であるリプライに着目する. Tweet の中には単語数が極端に少ないものがあり, 分析上ノイズとなることが考えられるため, 含まれる単語数が 4 以下の Tweet は除外した. 各分類器の訓練データは表 5.1 にまとめた. 各顔文字クラスタに属する Tweet の個数が全クラスタでほぼ同数になるようバランシングをしている. またクラスタ 12 のデータ数が少ないのは, 該当する Tweet が少なかったためである.

表 5.1. NaiveBayes 分類器の訓練データの詳細

	説明	総計 (5 クラス)	総計 (14 クラス)
多クラス分類	各クラスタ約 5 万 Tweet ずつ	約 25 万 Tweet	※約 67 万 Tweet
クラスタ別 2 クラス分類	対象クラスタ 5 万, 残りのクラスタからランダム抽出 5 万	約 10 万 Tweet	※約 10 万 Tweet
ポジネガの 2 クラス分類	各クラスタ約 5 万 Tweet ずつ	約 10 万 Tweet	無し
3 クラス分類	各クラスタ約 5 万 Tweet ずつ	約 15 万 Tweet	無し
	※クラスタ 12 のみ約 2 万 Tweet ずつ		

5.3　結果と考察

5.3.1　事前実験

事前実験として行った SVM について説明する [34]. SVM とは教師あり学習のアルゴリズムの 1 つであり, NaiveBayes と同様に Bag-of-Words モデルで特徴量ベクトルを作成し, 次元圧縮したものを訓練データに用いた. 次元圧縮とは高次元のデータを出来るだけ情報を保存するように低次元のデータに変換する事を意味し, 本研究では LSA(Latent Semantic Analysis) を用いて次元圧縮した. LSA は潜在意味解析と呼ばれるもので, 特異値分解による固有値を利用して次元圧縮している. SVM の実装にはオープンソースである libSVM を用い, 次元圧縮には scikit-learn ライブラリを用いた[*1]. 線形カーネルを用い, コストパラメータは 1 とし, 精度評価には NaiveBayes と同じく 10-fold cross-validation を用いて実験したところ, NaieveBayes 分類器の方が精度が良い事が確認された. そのため以降の実験は全て NaiveBayes 分類器を用いる.

5.3.2　顔文字ラベルを用いた感情推定

この節では, 顔文字クラスタをそのまま用いて多クラス分類を行う. 5 クラス分類及び 14 クラス分類のぞれぞれで実験した結果を表 5.2 にまとめた. ランダムゲッシングよりは高い正答率になっていることが分かる.

表 5.2. 5 クラス分類器及び 14 クラス分類器の Accuracy

	5 クラスタ分類	14 クラスタ分類
Accuracy	0.330	0.268

表 5.3 は 5 クラスタ分類の各クラスタについて 2 クラス分類の学習を行い, 分類器の精度をまとめたものである. クラスタ A, B, C, D については正解率が 6 割から 7 割で F-measure は 7 割前後となっている. またクラスタ E は正解率, F-measure 共に最も悪くなっている. これ

[*1] http://scikit-learn.org/stable/index.html

は前章で述べたようにクラスタ E は顔文字の傾向が最も掴みづらく雑多な集合であることに起因すると考えられる. 共起する単語の特徴が薄く, Bag-of-Words モデルのように単語の出現回数を特徴量とすると学習が難しいことを示唆している. また, 表 5.4 は 14 クラスタ分類の各クラスタについて 2 クラス分類の学習を行い, 分類器の精度をまとめたものである. 正答

表 5.3. 5 クラスタ分類の各クラスタの 2 クラス分類における評価値

クラスタ番号	A	B	C	D	E
Accuracy	0.635	0.657	0.614	0.640	0.572
Precision	0.637	0.671	0.627	0.666	0.604
Recall	0.693	0.787	0.725	0.745	0.729
F-measure	0.664	0.724	0.673	0.703	0.661

表 5.4. 14 クラスタ分類の各クラスタの 2 クラス分類における評価値

クラスタ	0	1	2	3	4	5	6
Accuracy	0.693	0.704	0.632	0.672	0.642	0.651	0.686
Precision	0.706	0.692	0.635	0.671	0.688	0.651	0.745
Recall	0.720	0.794	0.729	0.747	0.597	0.737	0.624
F-measure	0.713	0.740	0.679	0.707	0.640	0.691	0.679
クラスタ	7	8	9	10	11	12	13
Accuracy	0.675	0.660	0.650	0.654	0.678	0.838	0.661
Precision	0.665	0.654	0.652	0.643	0.677	0.816	0.649
Recall	0.774	0.764	0.726	0.780	0.752	0.641	0.785
F-measure	0.715	0.705	0.687	0.705	0.713	0.718	0.711

率が最も悪いのはクラスタ 2 であり, F-measure が最も悪いのはクラスタ 4 であることが分かる. これらはクラスタ E と同様に, 顔文字の傾向が掴みづらく雑多な集合であることに起因すると考えられる.

また多クラス分類の精度を下げている理由を調べるために, 5 クラス分類器において, 評価用データが正解クラスに対してどのクラスに推定されたか集計したマトリクスを作成した (表 5.5). 表 5.5 から「軽い Negative」であるクラスタ A に属するデータの誤分類先は「強い Negative」であるクラスタ B が多く, クラスタ B のデータの誤分類先はクラスタ A が多い事が分かる. また「強い Positive」であるクラスタ C に属するデータの誤分類先は「軽い Positive」であるクラスタ D が多く, クラスタ D のデータの誤分類先はクラスタ C が多い事が分かる. ここから Positive を表すクラスタ同士, 及び Negative を表すクラスタ同士が誤分類

表 5.5. 5 クラス分類器における分類結果をまとめたマトリクス

実際のクラスタ	A	B	C	D	E
A と推定された割合	0.406	0.292	0.180	0.171	0.225
B と推定された割合	0.241	0.353	0.114	0.103	0.126
C と推定された割合	0.133	0.134	0.288	0.224	0.215
D と推定された割合	0.118	0.140	0.295	0.388	0.225
E と推定された割合	0.102	0.080	0.122	0.114	0.208

されやすいと分かった．またクラスタ E はほぼ等しい割合で誤分類されており, これも雑多な集合であることに起因すると考えられる．この誤分類の結果を踏まえ, 次節では感情極性推定を行う．

5.3.3　顔文字ラベルを用いた感情極性推定

まず Tweet が Positive か Negative か判断する感情極性推定を行う．第 4 章で述べたように, 5 クラスタ分類におけるクラスタ A と B は Negative, クラスタ C と D は Positive となっており, また表 5.5 から分かるようにクラスタ A と B, 及び C と D は誤分類されやすく, 感情極性推定を行う場合はそれぞれ統合してよいと考えられる．クラスタ E の顔文字を Positive クラスや Negative クラスに統合すべきか検証する為, 以下の二つの分類器の精度を比較する．

分類器 1. 第 4 章の手法により顔文字を 2 クラスタに分類した結果を用いた 2 クラス分類器
分類器 2. 第 4 章の手法により顔文字を 2 クラスタに分類した結果から, クラスタ E に属する顔文字を除去した結果を用いた 2 クラス分類器

第 4 章の手法により顔文字を 2 クラスタに分類した結果と 5 クラスタに分類した結果において, 所属する顔文字の対応関係を表 5.6 にまとめた．2 クラスタ分類の結果はクラスタ A と B, 及び C と D が統合され, またクラスタ E が両方に分割された分類になっており, Positive を表す顔文字と Negative を表す顔文字に別れていると考えられる．分類器 1 と分類器 2 の精度を表 5.7 にまとめた．表 5.7 から, クラスタ E を Positive や Negative クラスに統合させると多少精度が落ちる事が分かる．
次にクラスタ E を Neutral と仮定し, Positive, Negative, Neutral に分ける 3 クラス分類器を作成する．顔文字のクラスタ E 以外の顔文字を Positive クラスと Negative クラスに統合する方法の違う三種類の分類器を作成する．

分類器 a. 前節の 5 クラス推定において, クラスタ A と B, 及び C と D 同士の誤分類を正答することで, Positive, Negative, Neutral の 3 クラス分類として評価値を算出する分類器

表 5.6. 2 クラスタ分類と 5 クラスタ分類の顔文字の所属の集計

クラスタ	0	1
クラスタ A	10	220
クラスタ B	4	114
クラスタ C	217	0
クラスタ D	203	2
クラスタ E	177	51

表 5.7. Positive クラス, Negative クラスの 2 クラス分類

	Acuracy		Precision	Recall	F-measure
全顔文字	0.671	Positive クラス	0.672	0.669	0.670
		Negative クラス	0.670	0.674	0.672
クラスタ E を除去	0.675	Positive クラス	0.676	0.673	0.674
		Negative クラス	0.674	0.677	0.676

分類器 b. 前節の 14 クラス推定において, クラスタ 1, 2, 8, 9, 11, 13 同士, クラスタ 0, 5, 7, 10 同士, クラスタ 3, 4, 6, 12 同士の誤分類を正答することで, Positive, Negative, Neutral の 3 クラス分類として評価値を算出する分類器

分類器 c. 第 4 章の手法により顔文字を 3 クラスタに分類した結果を用いた 3 クラス分類器

分類器 a は 5 クラス推定, 分類器 b は 14 クラス推定できるよう訓練し, 出力結果をまとめることで 3 クラス推定の精度として算出している. 分類 b では, 表 4.11 で言及したクラスタの対応関係を用いて, クラスタ A に対応するクラスタ 1, 9, 11 と B に対応する 8, 13 を Negative クラス, C に対応する 5, 7 と D に対応する 0, 10 を Positive クラス, E に対応する 3, 4, 12 を Neutral クラスとした. 対応関係が無いクラスタ 2 と 6 については, それぞれの顔文字が最も多く所属しているクラスタである A と E に統合した. 分類器 c で用いる, 第 4 章の手法により顔文字を 3 クラスタに分類した結果と 5 クラスタに分類した結果において, 所属する顔文字の対応関係を 5.8 にまとめた. 3 クラスタ分類の結果はクラスタ A と B, 及び C と D が統合された分類になっているため, Negative を表す顔文字, Positive を表す顔文字, 及び Neutral を表す顔文字に別れていると考えられる.

分類器 a, 分類器 b, 分類器 c の分類器の精度を表 5.9 にまとめた. より細かい顔文字分類の結果を用いた方が精度が良くなる事が分かる. またどの分類についても, Neutral クラスが最も指標値が低くなっている事が分かる.

表 5.8. 3 クラスタ分類と 5 クラスタ分類の顔文字の所属の集計

クラスタ	0	1	2
クラスタ A	2	6	222
クラスタ B	6	4	108
クラスタ C	26	191	0
クラスタ D	16	187	2
クラスタ E	224	0	4

表 5.9. Positive クラス, Negative クラス, Neutral クラスの 3 クラス分類

	Acuracy		Precision	Recall	F-measure
3 クラス分類	0.471	Positive クラス	0.452	0.532	0.489
		Negative クラス	0.506	0.604	0.551
		Neutral クラス	0.439	0.277	0.340
5 クラス分類を元にした 3 クラス分類	0.568	Positive クラス	0.555	0.557	0.556
		Negative クラス	0.568	0.581	0.574
		Neutral クラス	0.339	0.205	0.255
14 クラス分類を元にした 3 クラス分類	0.582	Positive クラス	0.526	0.552	0.539
		Negative クラス	0.643	0.698	0.669
		Neutral クラス	0.523	0.415	0.463

5.3.4　顔文字予測における特徴量の分析

この節では分類器の精度を改善するため, 入力する特徴量を変える事で精度を比較する. まず文中における単語の位置に着目し, 特徴量に組み込む単語の範囲を変え分類器を作成する. 次に品詞に着目し, 特徴量に組み込む単語の品詞別に分類器を作成する.

単語の位置別分析

これまでの実験ではテキスト中の単語を全て特徴量としていたが, テキストの中にはテキスト全体の表す感情と関係のない単語もある. 例えば「今さっき俺も見ました!キリショーほんといろいろ悩んだりしてたんですね (>_<)」のようなテキストにおいて, 前半はテキスト全体の感情に関係していない. ここで, 単語のテキスト全体の感情に対する影響力は文末 (文の区切り) からの距離に関係するという仮説を立てた. この項では, テキスト上のどの位置にある単語までがテキストの感情に影響しているのかを調査する.

図 5.1 では 5 クラス分類器について, 特徴量として考慮する単語を文末からの距離 (単語数) が n 以内の単語とし,n の値を変化させ精度をまとめた結果を, 全単語を考慮した場合と比較し

ている．横軸の値が考慮した最大距離, 縦軸が推薦精度となっている．例えば最大距離を 5 とする場合,

キリ— ショー ほんと いろいろ 悩ん で たんす ね (>_<)

という文のうち「いろいろ」,「悩ん」,「で」,「たんす」,「ね」の 5 つを特徴量として入力する．

この図から, 考慮する単語の範囲が短すぎるとテキストの感情を推定出来ないが, 逆に範囲が広すぎても推薦精度が落ちており, 5 単語付近で最も精度が高い事が分かる．

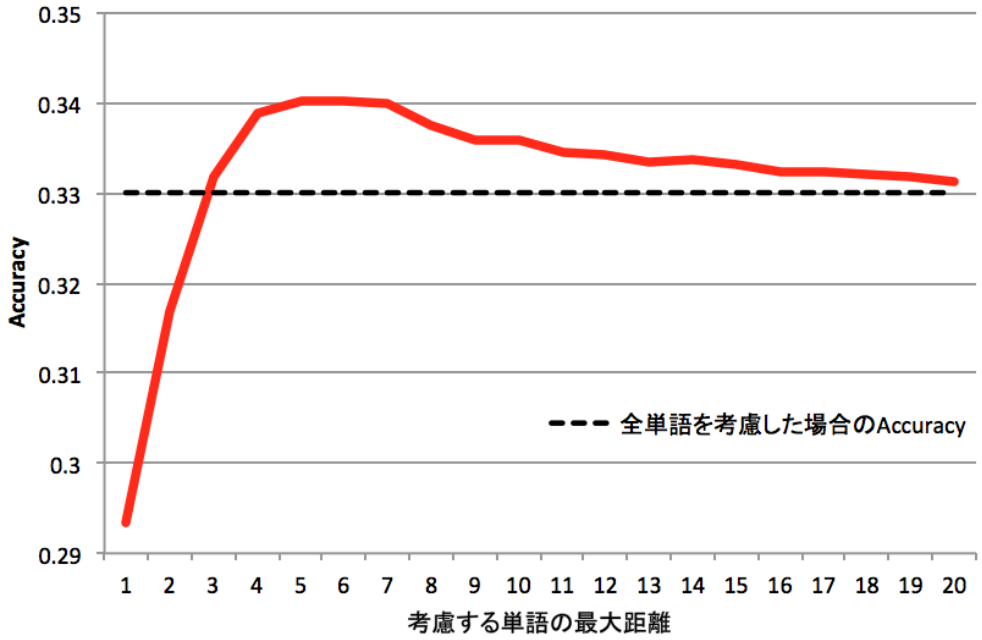

図 5.1. 考慮する単語の最大距離に対する 5 クラス分類器の精度

図 5.2 に 14 クラス分類について同様の実験を行った結果を記した．この表から, 14 クラス分類においても全単語を考慮するより最大距離を 14 単語付近とした場合が最も精度が上がる事が分かる．しかし, 全単語を考慮した場合と比べて精度の向上はわずかとなっている．

以上の結果から, 感情推定を行う際は全単語を考慮するのではなく, 単語の位置により足切りをした方が良い推定が出来る事を示唆している．また 14 クラス分類の精度が最大となる単語距離は 5 クラス分類と比べて大きく, 精度の向上が小さかった理由は．14 クラス分類はより細かい分類を行おうとしているため, 似たようなクラスを正しく分類する為にはより遠い単語の情報も必要になるからだと考えられる．

品詞別分析

感情分析に際し, 品詞 (part-of-speech) を特徴として考慮する論文は多く有り, 本研究でも品詞について言及することにする．この項ではどの品詞に属する単語群が Tweet の感情に影響するのかを調べる為に, 分類器に入力する特徴量を品詞別にすることで精度の変化を比較す

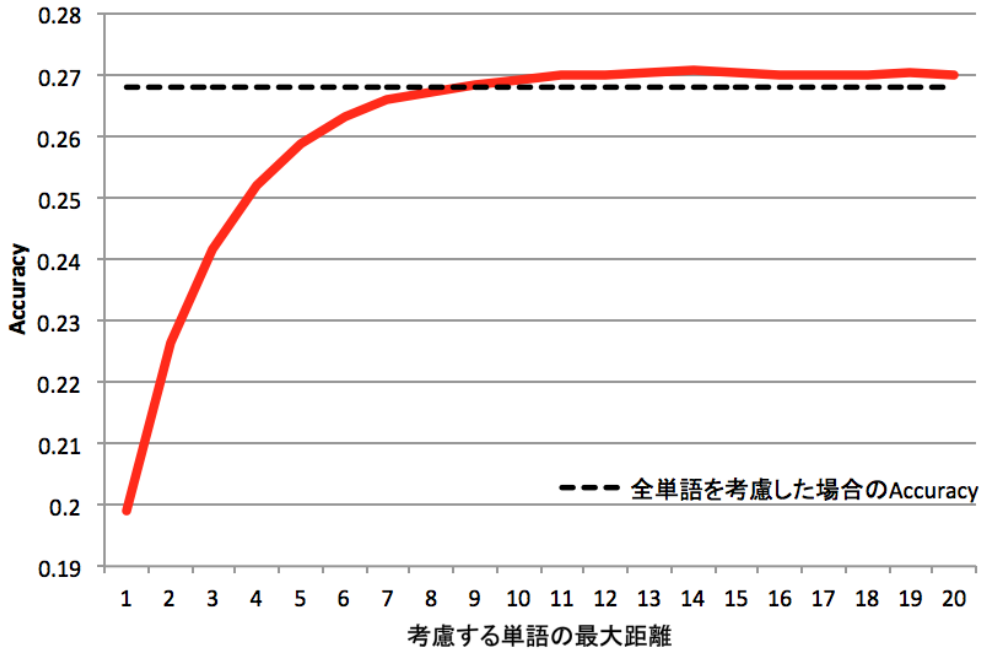

図 5.2. 考慮する単語の最大距離に対する 14 クラス分類器の精度

る. 調べる品詞は [名詞], [動詞], [形容詞, 副詞], [未知語], [名詞, 動詞, 形容詞, 副詞], [名詞, 動詞, 形容詞, 副詞, 未知語] とする. ここでは分類器は 5 クラス分類器とし, 単語はすべて特徴語として用いる.

図 5.3 は横軸に記す各品詞を特徴語とした場合の推薦精度をまとめている. 図 5.3 から, 一つの品詞に限る場合は名詞を用いると最も精度が良くなる事が分かる. またどの品詞も特徴量に組み込む事で精度が落ちる事は無く, 感情を判断する因子となっている事が分かる. 名詞が最も精度が高くなる理由の一つとしての単語の出現数の違いが挙げられる. 表 5.10 は訓練データにおける各品詞単語の出現割合を集計したものである. ただし, 用いたデータは節 5.3.2 における 5 クラス分類器について, `K-fold Cross-validation` で分割されている 10 種類の訓練データ群の内の一つである. 名詞が半数を占めており, 次に多いのが動詞となっていることが分かる. これは図 5.3 の精度の順位と一致しており, 名詞や動詞を用いた場合の精度が高かったのは, 単語数が多いため良く学習する事ができたからだと考えられる. また参考としてどのような名詞が影響しているのか確認するため, 表 5.11 に 4 章と同じ方法で分類器に入力された単語を指標値 $R_j(i)$ の高い順に並べた. ここで入力しているデータは, 上記の名詞分類器の内, `K-fold Cross-validation` で分割されている 10 種類の訓練データ群の内の一つを用いている.

この表によると, 「最悪」, 「残念」, 「素敵」, 「綺麗」等, 形容動詞の単語が形態素解析により語幹を名詞として抽出されたものが多いように思える.

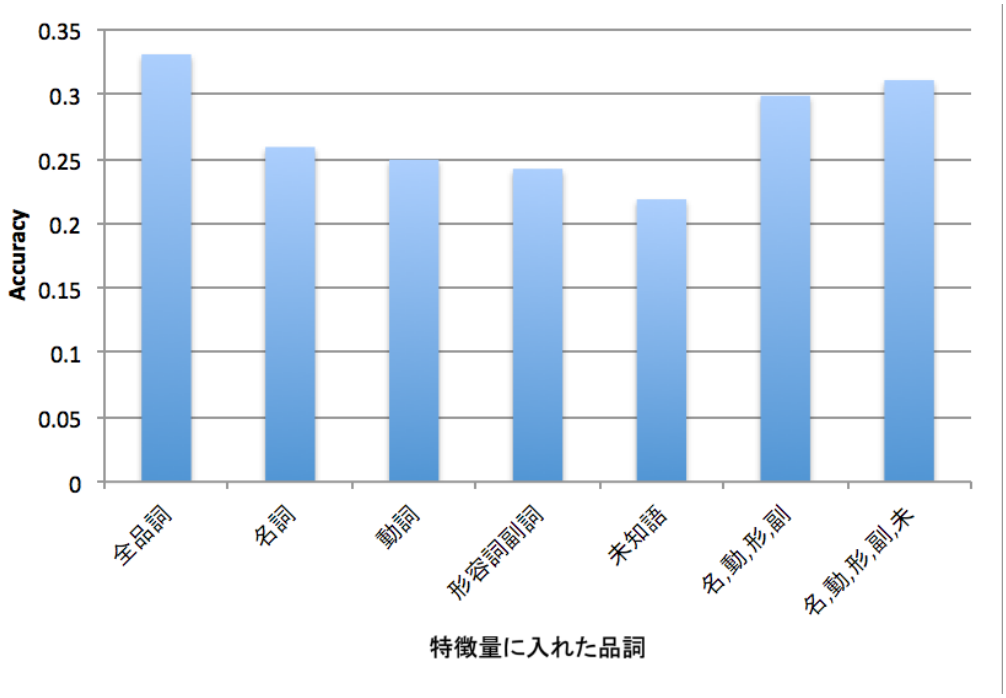

図 5.3. 品詞別の分類器の精度

表 5.10. Tweet に登場する品詞の割合

名詞	動詞	形容詞	副詞	未知語	その他
0.504	0.195	0.075	0.056	0.0546	0.116

表 5.11. 名詞の特徴語 ($R_j(i)$)

A	微妙 最悪 急 あら 不安 ダメ ギリギリ 性 勘違い しょうが 記憶 ネット 苦手 変 大変 謎 あんま ビックリ 夜中 嫌 社会 会社 後悔 生活 天気 内容 可能 迷惑 体調 意外 あそこ 普段 雨 結局 確か 状態 地震 最近 へん カード くせ 仕方 夕方 タイミング ひと 雪 あいつ 風邪 残念 緊張
B	ショック 涙 残念 感動 申し訳 地震 心配 不安 体調 風邪 無理 病院 大事 大変 実習 試験 苦手 嫌 緊張 現実 最後 雨 最悪 ダメ 課題 熱 本当 途中 微妙 時期 返事 薬 だめ 返信 ホント お金 ほんま 我慢 携帯 調子 いや 交換 チケット 週間 クラス や タイミング すぎ うそ 気持ち
C	最高 素敵 しま 楽しみ 誕生 幸せ 是非 歓迎 大好き アイコ 気軽 め 祭り あざ 綺麗 久しぶり 天使 げ 翔 もも カッコ 仲間 遊び 笑顔 タメ ぶり 好き しょ 観 すき ン おいで テンション 会話 爆笑 安定 ほか 春 かお 妄想 神 なっちゃん こちら 雰囲気 世界 メンバー フォロー めちゃくちゃ 今度 ドキドキ
D	了解 フォロー こちら 気軽 お願い おいで 歓迎 お気 報告 鍋 タメ えり 今度 是非 楽しみ 語 お世話 かい あざ ほか 紹介 ごはん 久しぶり 大切 疲れ 確認 がん 遊び 元気 いつか 屋 サッカー 機会 ドラマ いち 応援 感想 アルバム 人生 十 綺麗 無事 姉 さい 安心 連絡 来週 すき よ 感謝
E	乙 メモ 三 待機 がん お菓子 エロ 用 あなた 全力 期 9 けい 本気 おじさん ばか 当日 報告 写 後ろ 適当 お互い 変態 ゲーム ろ 0 翔 結婚 勉強 想像 応援 基本 間違い 彼女 チェック 愛 汗 犬 ラーメン 勝手 水 弟 神 顔 なか ゆか 女子 絵 大会 半分

第 6 章

結論

本研究は日本語テキストで利用されている多種多様な顔文字を分類をテキストの感情推定に利用するという着想から, ソーシャルメディアにおける日本語テキストの顔文字を対象に, 顔文字の自動分類による感情ラベルに資する情報の抽出, および顔文字を利用した短文テキストの感情推定の手法の構築を目的として行った.

研究の前半では, 顔文字をテキストのラベルとして用いるために, 大規模なユーザの書き込みデータから多数の顔文字の自動分類とラベル表現の抽出を行った. まず分散表現モデルを用いて顔文字が使われるコンテクストをベクトルで表現し, クラスタリングすることで分類した. 次にユーザの書き込みデータにおいて顔文字と共に出現する単語の一覧を比較することで顔文字ラベル表現を抽出した. 分類結果は既存の研究結果とも親和性のある性質が確認され, また単語の抽出においては, TFIDF やそれを改善した指標値を用いる事で各顔文字クラスタの特徴を表す感情語を抽出することができた. TFIDF よりも改善した指標値を用いた方が端的に感情を表す単語が多く抽出された. これは一般に文書の特徴を表す際は TFIDF が用いられるが, 感情ラベルに資する感情語を抽出するというタスクにおいては, 抽出基準を工夫すべきであることを示唆している. また Twitter のデータにおける 1000 個の顔文字を対象としたが, 本手法は大規模なデータ上の膨大な顔文字の分類に対応できるところに意義があり, 他のソーシャルメディアへの応用やリアルタイムのデータを用いて情報の更新が可能である. 本研究では顔文字をラベルとした教師あり学習を行うために分類を行ったため, 主要な目的は分類する事自体にあり, 感情ラベルに関しては共に使われている単語群からのラベル表現の抽出程度とした. しかし本手法による顔文字分類を, 明確な感情ラベルが必要なタスクに用いる場合は感情を定量的に示す手法が望ましいと考えられ, 今後の研究が待たれる.

研究の後半では本手法による顔文字の分類結果を用いて, 顔文字を含むユーザの書き込みデータをラベル付きテキストと見なす事で訓練データとし, 教師あり学習による感情推定を行った. 顔文字ラベルを利用した多クラス感情推定, 及び Positive, Negative, Neutral の感情極性推定のベースラインを示した. 雑多な顔文字の集合を Neutral としたため, Neutral クラスの精度が悪く, 今後の課題として雑多な顔文字グループの処理を考えなくてはならないだろう. また作成した分類器を用いて, 感情推定に影響する特徴量を分析した. テキストの感情に影響する単語の位置を分析したところ, クラスタ数が少ない場合には文の区切りから一定の距

離内の単語を考慮すると良い事が分かった．またテキストの感情に寄与する品詞を分析したところ, 名詞が最も寄与していることが分かった．これはデータの中で名詞の出現割合が最も高い事に起因すると考えられる．本研究では, 提案する手法のベースラインの確認に焦点を当てたため, 精度の向上には改善の余地がある．今後の展開として, NaiveBayes 以外のアルゴリズムによる分類器の作成や, 特徴量の設計の変更の精度の改善が望まれる．特に, Bag-of-Words モデルでは単語の出現回数のみを考慮しており, 単語の順番や意味関係を全く考慮していないため, 従来研究が示しているように Unigram だけでなく高次の n-gram を考慮した Bag-of-n-grams モデルを用いることや, 分散表現モデルによるコンテクストの情報を含む特徴量を用いる事も考えられる．本研究の手法を用いた感情推定の拡張や改善となる研究がなされれば, 筆者の幸いとするところである．

謝辞

本研究を行うにあたり, 東京大学大学院工学系研究科 森純一郎 特任講師にはテーマ選びから関連論文の探し方などの基礎的なところ, さらに研究の具体的なアドバイスや論文を書く姿勢や構成・執筆に至るまで丁寧なご指導ありがとうございました. また, 東京大学大学院工学系研究科 坂田一郎教授には論文投稿に関わる様々なご調整や研究のアドバイス等ご指導ありがとうございました. 株式会社ホットリンク R&D 部門統括かつ東京大学大学院工学系研究科客員研究員 榊氏には, 通常では手に入らない貴重なデータのご提供他, 構成・執筆に至るまで丁寧なご指導ありがとうございました. 東京大学大学院工学系研究科 特任研究員 原氏 には研究の指針や具体的なアドバイスをして下さり, 心から感謝しています. 森研究室の博士課程 3 年の丸井氏には, 研究の指針をいただいたただけではなく, 度々研究の具体的な相談に乗って頂き, さらには論文執筆の丁寧なご指導を頂き, 大変お世話になりました. 研究室の先輩方, 坂田研究室修士課程 2 年の早嶋氏, 沢村氏, 森研究室修士課程 1 年の山下氏, 株田氏にはいつも優しく接していただき, また時には研究についてアドバイスをいただき, 思い出深い研究生活を送ることができました. 学部 4 年の石塚君・小林君・上子さん・サンタモンさんは, 最高の仲間でした. 改めて皆様に深く感謝すると共に, 今後のご多幸をお祈り申し上げます.

参考文献

[1] O'Connor, Brendan, et al. "From tweets to polls: Linking text sentiment to public opinion time series." *ICWSM* ,11 (2010): 122-129.

[2] Ghiassi, M., J. Skinner, and D. Zimbra. "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network."*Expert Systems with Applications: An International Journal*, 40.16 (2013): 6266-6282.

[3] Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." *CS224N Project Report, Stanford*, (2009): 1-12.

[4] Gajadhar, J., and Green, J. The importance of nonverbal elements in online chat. , *Educause Quarterly*, 28.4 (2005):2.

[5] Walther, J. B., and D' addario, K. P. The impacts of emoticons on message interpretation in computer-mediated communication. *Social Science Computer Review*, 19.3 (2001):324 347.

[6] Derks, D.; Bos, A. E. R.; and Grumbkow, J. "Emoticons and social interaction on the Internet: the importance of social context", *Computers in Human Behavior*, 23 (2007):842849.

[7] J. Park, V. Barash, C. Fink, and M. Cha. Emoti- con Style: Interpreting Differences in Emoticons Across Cultures. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, (2013) 466475. AAAI Press.

[8] Yuki, M.; Maddux, W. W.; and Masuda, T. Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States. *Journal of Experimental Social Psychology*, 43.2 (2007):303311.

[9] Markman, K. M., and Oshima, S. Pragmatic play? some possible functions of english emoticons and japanese kaomoji in computer-mediated discourse. In *Association of Internet Researchers Annual Conference.* (2007)

[10] Tanaka, Yuki, Hiroya Takamura, and Manabu Okumura. "Extraction and classification of facemarks."*Proceedings of the 10th international conference on Intelligent user interfaces.*, (2005) ACM.

[11] Ptaszynski, Michal, et al. "Towards Fully Automatic Emoticon Analysis System

(o)."*Proceedings of The Fifteenth Annual Meeting of The Association for Natural Language Processing (NLP-2010), Tokyo.*, (2010)

[12] 山本湧輝, 熊本忠彦, and 灘本明代. "Twitter 特有表現を考慮したツイートの多次元感情抽出手法の提案."情報処理学会関西支部支部大会講演論文集, (2014): 5p.

[13] 川上正浩. "顔文字が表す感情と強調に関するデータベース." 大阪樟蔭女子大学人間科学研究紀要 7 (2008): 67-82.

[14] 江村優花, and 関洋平. "テキストに現れる感情, コミュニケーション, 動作タイプの推定に基づく顔文字の推薦."情報処理学会研究報告. *DD,[デジタル・ドキュメント]*, 2012.1 (2012): 1-7.

[15] 吉田綾奈, 邱起仁, and 櫨山淳雄. "顔文字推薦のための感情を付与した顔文字データベースの構築." 情報処理学会研究報告. *HCI,* ヒューマンコンピュータインタラクション研究会報告, 2014.35 (2014): 1-6.

[16] 原田登美. "「顔文字」 による日本語の円滑なコミュニケーション:「配慮」 と 「ポライトネス」 の表現機能."言語と文化, 8 (2004): 205-224.

[17] Turney, Peter D. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews."*Proceedings of the 40th annual meeting on association for computational linguistics.*, (2002) Association for Computational Linguistics.

[18] Agarwal, Apoorv, Fadi Biadsy, and Kathleen R. Mckeown. "Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams."*Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.*, (2009) Association for Computational Linguistics.

[19] Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining."*LREC.*,(2010) Vol. 10.

[20] Thelwall, Mike, Kevan Buckley, and Georgios Paltoglou. "Sentiment strength detection for the social web." *Journal of the American Society for Information Science and Technology*, 63.1 (2012): 163-173.

[21] Thelwall, Mike, and Kevan Buckley. "Topic‐based sentiment analysis for the social web: The role of mood and issue‐related words." *Journal of the American Society for Information Science and Technology*, 64.8 (2013): 1608-1617.

[22] Bravo-Marquez, Felipe, Marcelo Mendoza, and Barbara Poblete. "Combining strengths, emotions and polarities for boosting Twitter sentiment analysis."*Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining.*, (2013)ACM.

[23] Read, Jonathon. "Using emoticons to reduce dependency in machine learning techniques for sentiment classification."*Proceedings of the ACL Student Research Workshop.*, (2005) Association for Computational Linguistics.

[24] Kouloumpis, Efthymios, Theresa Wilson, and Johanna Moore. "Twitter sentiment

analysis: The good the bad and the omg!."*ICWSM* 11 (2011): 538-541.

[25] Purver, Matthew, and Stuart Battersby. "Experimenting with distant supervision for emotion classification." *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics.*, (2012) Association for Computational Linguistics.

[26] Suttles, Jared, and Nancy Ide. "Distant supervision for emotion classification with discrete binary values."*Computational Linguistics and Intelligent Text Processing.*,(2013) : 121-136, Springer Berlin Heidelberg,

[27] 風間 一洋 （和歌山大学）, 榊 剛史 （東京大学）, 鳥海 不二夫（東京大学）, 篠田 孝祐（慶應義塾大学/理化学研究所）, 栗原 聡 （電気通信大学）, 野田 五十樹（産業技術総合研究所）. " 顔文字に着目したツイートの感情変化の分析." *WebDB Forum2013*

[28] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality."*Advances in Neural Information Processing Systems.*, (2013).

[29] MacQueen, J. B. "Some Methods for classification and Analysis of Multivariate Observations".*In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability.*, (1967) pp. 281297.

[30] Davies, David L., and Donald W. Bouldin. "A cluster separation measure." *Pattern Analysis and Machine Intelligence, IEEE Transactions on 2* ,(1979): 224-227.

[31] 工藤拓, 山本薫, and 松本裕治. "Conditional Random Fields を用いた日本語形態素解析." 情処学 NL 研報 (2004): 161-13.

[32] Jones KS "A statistical interpretation of term specificity and its application in retrieval". *Journal of Documentation.*, 28.1 (1972): 1121,

[33] Jurafsky, Daniel, and H. James. "Speech and language processing an introduction to natural language processing, computational linguistics, and speech." 17 (2000):636

[34] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." *Machine learning*, 20.3 (1995): 273-297.