ツイート中の主観極性同定モデルの訓練事例収集における絵文字の利用†

林 友超*¹・崔 鴻翌*¹・宇津呂 武仁*²

本論文では、ツイート中の主観極性同定モデルの訓練事例の作成にあたって、絵文字の有無および絵文字の種類を手がかりとして自動収集した訓練用ツイートを用いてツイート中の主観極性同定モデルを訓練する方式を提案する。そして、主観極性を人手で同定した訓練事例との間で性能比較を行い、人手を介さず自動収集した 9,000～136,212 ツイートの持つ性能が、主観極性を人手で同定した 270～540 ツイートの持つ性能に匹敵することを示す。

キーワード：ツイッター、絵文字、主観極性同定

1. はじめに

近年、ツイッター等の SNS が普及してきた。ツイッターの利用者の多くは、自身が配信するツイートの中に、様々な事柄に対する主観情報を含めることがよく知られている。ツイッターでは、文字数制限があるため、制限された文字数の範囲で適切に自分の感情を表すことが難しい。そこで、ツイッターでは、制限された文字数の範囲で感情を適切に表現するために、絵文字が利用される傾向がある。それに対して、自然言語処理の研究分野においては、絵文字を含めたテキストの分析手法が研究対象の 1 つとして注目されている。

本論文では、ツイート中の主観を自動分析することを目的とする。特に、本論文では、ツイート中において絵文字以外のテキスト部分が表す主観極性に着目し、「ツイート中の絵文字以外のテキスト部分」を入力して、肯定・否定・中立といった、「ツイート中の絵文字以外のテキスト部分」が持つ主観極性を同定するモデルを「主観極性同定モデル」と呼ぶ。通常、主観極性同定モデルを訓練する際には、人手で予め主観極性同定済みの訓練・開発事例集合を用意し、これを教師データとして主観極性同定モデルを訓練するのが一般的である。しかし、このアプローチでは、訓練・開発事例集合の作成コストが高いため、大規模訓練事例を作成することは容易ではない。そこで、本論文では、ツイート中の主観極性同定モデルの訓練事例の作成にあたって、絵文字の有無および絵文字の種類を手がかりとして、人手を介さず自動で訓練用ツイートを収集し、これを用いてツイート中の主観極性同定モデルを訓練する方式を提案し、その性能の評価結果を報告する。

本論文では、まず、ツイートが絵文字を含むか否か、また、絵文字を含む場合には、主観を表す特定の絵文字を含むか否かに着目する。具体的には、「ハッピー」を表す絵文字を含むツイート、および、「怒る」を表す絵文字を含むツイートのテキスト部分に含まれる主観の種類・内容を人手で分析した。その結果、最大約 41%（表 2 に示す詳細の通り、約 17～41%）程度の割合で、絵文字が表す主観とテキスト部分が表す主観が矛盾するツイートが存在していた。¹一方、絵文字を含まないツイートのテキスト部分の主観を分析した結果においては、約 20% 程度の割合で、テキスト部分に主観が含まれるツイートが存在していた。以上の分析については、次節でその詳細を示す。この分析結果をふまえて、本論文では、「ハッピー」を表す絵文字が肯定的な主観極性を、「怒る」を表す絵文字が否定的な主観極性を、それぞれ表すとして、各絵文字の表す主観極性をそのままツイート自身の極性とみなすアプローチを採用する。同様に、絵文字を含まないツイートに対しては、そのツイート自身が中立的な主観極性を持つとするアプローチを採用する。そして、「ハッピー」を表す絵文字を含むツイート、「怒る」を表す絵文字を含むツイート、および、絵文字を含まないツイートをそれぞれ自動収集し、これらをツイート中の主観極性同定モデル訓練事例として用いて、ツイートに対する肯定・否定・中立の主観極性を同定するモデルを訓練する。その一方で、本論文のアプローチがどの程度の性能を達成するのかを検証するため、人手で肯定・否定・中立極性を同定したツイート集合を収集し、これを訓練事例としてツイートに対する肯定・否定・中立の主観極性を同定するモデルを訓練し、本論文のアプローチによって訓練した主観極性同定モデルとの間で性能の比較を行う。

本論文では、ツイート中の主観極性同定モデルとしては、畳み込みニューラルネットワーク（convolutional neural network; CNN）および SVM（support vector machines）を用いる。この

† Utilizing Emoji in Collecting Training Instances of a Model for Sentiment Analysis of Tweets
Youchao LIN, Hongyi CUI, and Takehito UTSURO

*¹ 筑波大学大学院システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba

*² 筑波大学システム情報系
Faculty of Engineering, Information and Systems, University of Tsukuba

1 絵文字が表す主観の種類としては、「ハッピー」および「怒り」以外にも多様なものが存在するが、他の多様な主観を表す絵文字と比較して、ツイートのテキスト部分が表す主観との相関が相対的に高い傾向にある主観の種類として、本論文では「ハッピー」および「怒り」の主観を対象とする。

うち、CNN は、自然言語処理における文分類タスクにおいて高い性能を示すことが報告されている [1–3]。例えば、文献 [2] においては、訓練済みの分散表現を用いる文分類手法により高い性能を達成している。これらをふまえて、本論文では、絵文字付きツイートを訓練事例として、CNN および SVM の訓練を行う手法を提案する。

結果として、本論文では、評価実験において、CNN による主観極性分類性能の方が SVM よりも優れていることを示す。また、絵文字の有無に基づき主観極性を自動同定したツイートを用いて訓練した主観極性同定モデル M_{auto} に対して、人手で主観極性を同定したツイートを用いて訓練した主観極性同定モデル M_{man} との間で性能の比較評価を行った結果を示す。その結果においては、絵文字を含むツイート、および、絵文字を含まないツイート、合計 9,000~136,212 ツイートを自動収集したものを訓練・開発事例として訓練した主観極性同定モデルの性能が、主観極性を人手で同定した訓練・開発事例 270~540 ツイートを用いて訓練した主観極性同定モデルの性能に匹敵することが分かった。この評価結果をふまえて、本論文では、絵文字を含むツイート、および、絵文字を含まないツイートを自動収集して訓練・開発事例として用いるアプローチが有効であることを示す。

以下、2 節においては、絵文字を含むツイート、および、絵文字を含まないツイートにおける主観極性の分析を行った結果について述べる。3 節においては、本論文におけるツイートの主観極性同定タスクに対して適用する CNN モデルについて述べる。4 節においては、本論文の評価実験において共通に用いる評価事例について述べる。5 節、および、6 節においては、モデル訓練用ツイート事例における主観極性判定を人手で行った場合、および、絵文字の有無に基づき自動で行った場合の各々について、評価結果の分析を行う。

2. 絵文字を含むツイートにおける主観の人手分析

本論文において、「ハッピー」を表す絵文字、「怒る」を表す絵文字を含むツイート、および、絵文字を含まないツイートのテキスト部分に含まれる主観の種類・内容を人手で分析したところ、最大約 41%（本節の具体的な調査結果では、表 2 に示す通り、約 17~41%）程度の割合で、絵文字が表す主観とテキスト部分が表す主観が矛盾するツイートが混在していた。本節では、この人手分析の詳細について述べる。

まず、2016 年 11 月 13 日から 2017 年 6 月 23 日までの約 7 ヶ月の期間において、Twitter Streaming API² において filter として特定のキーワードを指定せずに収集した日本語ツイート³の総数を、表 1 の「総数」欄に示す。次に、本論文において

を、「怒り」を表す絵文字として

表 1 Twitter Streaming API によって収集したツイートにおける絵文字含有率

		ツイート数	割合(%)
絵文字を含むツイート	「ハッピー」	50,627	0.60
	「怒る」	2,926	0.03
		1,085	0.01
絵文字を含まないツイート		7,597,377	90.39
総数		8,405,298	100.00

を、それぞれ収集・分析対象として用いる。⁴ 表 1 においては、収集結果のツイート集合中において、これらの各絵文字を含むツイートの数とその割合を「絵文字を含むツイート」欄に示す。同様に、これら 3 種類の絵文字だけでなく主観を表し得る絵文字およびその他の絵文字を含まないツイートの数とその割合を「絵文字を含まないツイート」欄に示す。⁵

ここで、Twitter Streaming API において filter として特定のキーワードを指定せずツイートを収集する方法では、表 1 に示すように、「ハッピー」と「怒る」の絵文字の間で収集できるツイート数の比率が極めて偏っており、かつ、「怒る」の絵文字を含むツイート数の収集効率も相対的によくない。そこで、本論文では、収集・分析対象とする全ての絵文字に対してツイート収集条件を揃えとともに、相対的に規模の大きいツイート集合をより効率よく収集するために、収集・分析対象とする全ての絵文字に対して、別途、2016 年 11 月 13 日から 2017 年 6 月 23 日までの約 7 ヶ月の期間において、Twitter Streaming API において各絵文字を filter として指定して日本語ツイートを収集し⁶、分析対象および訓練・開発・評価用事例の情報源とする。この手順によって収集されたツイートの総数を、表 2 の「3 種類の絵文字のうち、1 種類のみを含み、他の 2 種類の絵文字を含まないツイート」の項目の「収集総数」欄に示す。

次に、本論文では、各絵文字を含むツイート、および、絵文字を含まないツイートのテキスト部分が表す主観の分析を行う。具体的には、絵文字を含むツイートとしては、表 2「収集

4 「ハッピー」および「怒り」を表す絵文字としては何種類かのものが存在する。そのうち、本論文では、「ハッピー」を表す絵文字としては、 が最も高頻度であり、かつ、 1 種類だけで十分な数のツイートを容易に収集可能であるので、 1 種類のみを収集・分析対象とする。一方、「怒り」を表す絵文字は相対的に低頻度であり、1 種類の絵文字を対象とするだけでは十分な数のツイートの収集が容易ではなかったため、最も高頻度な 2 種類の絵文字 および を収集・分析対象とする。

5 厳密には、ここでは、絵文字全体の約 79%（約 1,000 種類）を含まないツイートを対象としている。この約 1,000 種類の中には、主観を表し得る絵文字は全て含まれるが、それ以外の主観とは無関係な絵文字も多く含まれる。

6 本論文で収集・分析対象とした絵文字 、 および のうちの 2 種類以上の絵文字を含むツイートについては、そのテキスト部分の主観がツイート中のどの絵文字の主観に従うのかが自明ではないため、本論文では、ツイートとしては、これら 3 種類の絵文字のうち 1 種類のみを含み、その他の 2 種類の絵文字を含まないものを収集対象とする。また、これら 3 種類以外にも多様な主観を表す絵文字が存在し、ツイート中にそれらが含まれる可能性も考えられるが、それらの頻度は相対的には小さいとみなして、本論文では、これら 3 種類以外の絵文字が含まれるツイートを除外することはしなかった。

2 <https://dev.twitter.com/streaming/overview>

3 特定のキーワードを指定せず収集を行う場合、リアルタイムで流れているツイートから無作為に一定数のツイート集合が収集される。

表2 絵文字を含むツイート・絵文字を含まないツイートにおける主観の分類

三種類の絵文字のうち、一種類のみを含み、他の二種類の絵文字を含まないツイート							
		収集総数	調査総数	テキスト部分と絵文字の主観が合致		テキスト部分と絵文字の主観が矛盾	
				ツイート数	割合(%)	ツイート数	割合(%)
ハッピー	😊	387,412	120	100	83.33	20	16.67
怒る	😡	29,765	85	50	58.82	35	41.18
	😠	15,640	70	50	71.43	20	28.57
絵文字を含まないツイート							
		収集総数	調査総数	ツイートの主観極性なし		ツイートの主観極性あり	
				ツイート数	割合(%)	ツイート数	割合(%)
		7,597,377	122	100	81.97	22	18.03

総数」欄のツイートを対象として、

- ツイート中のテキスト部分が表す主観と絵文字が表す主観が**合致**（例: 😊 私はとても幸せです！）
- ツイート中のテキスト部分が表す主観と絵文字が表す主観が**矛盾**（例: 😡 仕事を失った！）

の2種類への分類を行った。一方、絵文字を含まないツイートとしては、表2「収集総数」欄のツイート（表1における「絵文字を含まないツイート」と同一）を対象として、

- ツイート中のテキスト部分が肯定極性または否定極性の主観を含む
- ツイート中のテキスト部分が肯定極性または否定極性の主観を含まない

の2種類への分類を行った。⁷分類結果の統計を表2に示す。このうち、特に、次に、「テキスト部分と絵文字の主観が矛盾」となるツイートのテキスト部分を人手で分析したところ、おおよそ、

- 「主観と絵文字が逆」—テキスト部分の主観と絵文字の主観が逆の極性となる。
- 「主観なし」—テキスト部分には主観が含まれない（ただし、「主観なしの広告」以外）。
- 「主観なしの広告」—テキスト部分には主観が含まれず、かつ、その内容は広告である。
- 「日本語ではない」—日本語以外の言語で書かれたツイートである。

の4種類に分類された。これらの内訳および例を表3に示す。また、絵文字を含まないツイートにおいて、主観を含むツイートのテキスト部分の主観を人手で分析したところ、肯定極性の割合が81.8%、否定極性の割合が18.2%であった。

3. CNNを用いた文の分類

本論文では、日本語ツイートの主観極性同定タスクに対して、CNNモデルによる文分類手法[2]を適用する。

7 絵文字😊を含むツイートについては、「テキスト部分と絵文字の主観が合致」となるツイートの数が100になるまで分類対象となるサンプルを抽出した。また、絵文字😡、または、😠を含むツイートについては、それぞれ、「テキスト部分と絵文字の主観が合致」となるツイートの数が50になるまで分類対象となるサンプルを抽出した。一方、絵文字を含まないツイートについては、「テキスト部分の主観が中立」となるツイートの数が100になるまで分類対象となるサンプルを抽出した。

3.1 CNNモデルの構成

文分類手法[2]におけるCNNモデルのアーキテクチャを図1に示す。以下では、文分類手法[2]におけるCNNモデルの設定の詳細について述べる。本論文では、TensorFlow⁸上のCNNパッケージを用いてCNNモデルの実装を行う。

まず、入力文を行列に変換する。具体的には、入力文が n 形態素から構成され、入力文の i 番目の形態素 $x_i (\in R^k)$ が k 次元のベクトルで表されるとすると、入力文の行列の次元数は $n \times k$ となる。ただし、実際には、入力文集合中で形態素数が最大値 $L (\geq n)$ となる最長文（形態素数最大文）にあわせて、全ての入力文において固定サイズ $L \times k$ の次元数の行列を用い、形態素の存在しない $(L - n) \times k$ 次元分の要素は0で埋める（ゼロパディング）。次に、畳み込み層では、 i 番目から $i + h - 1$ 番目の形態素に対して、次式に従ってサイズ h のフィルタ行列 w が適用され、特徴 c_i が生成される。

$$c_i = f(w \cdot x_{i:i+h-1} + b) \quad (1)$$

ここで、 f は活性化関数（ReLUを用いる）であり、 b はバイアスである。入力文に対して、各フィルタサイズごとにこの過程を適用し、生成された特徴マップに対して、max-over-timeプーリング関数を適用し、固定長ベクトルを生成する。次に、正則化のため、最後から2つ目の層でドロップアウト関数を適用する。最後に、各特徴マップから得られた固定長出力を全結合層に通し、softmax関数を適用することにより、各ラベルについての確率分布が得られる。パラメータ最適化の過程においては、確率勾配降下法（SGD）と誤差逆伝播法を用いる。

フィルタサイズ h を3, 4, 5に設定し、それぞれのフィルタのサイズの出力特徴マップ数を100に固定する。ミニバッチサイズを50, epoch数の上限を50とし、パラメータ最適化誤差軽減法としてAdam[4]⁹を用いる。

3.2 単語分散表現の設定

本論文では、ツイートに対して単語分散表現を求めるにあたって、まず、ツイート中の絵文字¹⁰、ユーザー名（例: “@xxx”

⁸ <https://www.tensorflow.org/>

⁹ Kim[2]においては、パラメータ最適化誤差軽減法としてADADELTA[5]が用いられているが、本論文では、ADADELTAよりも性能がよいことが報告されているAdam[4]を用いる。

¹⁰ 厳密には、2節で分析対象とした3種類の絵文字、および、その他の主観を表し得る全ての絵文字を含む約79%の絵文字（具体的には、

表3 「絵文字を含むツイート」中のテキスト部分の主観と絵文字の主観が矛盾する場合の内訳、および、「絵文字を含まないツイート」中のテキスト部分の主観極性の内訳

		分類	例	数(割合(%))	
絵文字を含む ツイート	ハッピー😊	テキスト部分と 絵文字の主観が逆	朝からうざいって 😊	6 (30.0)	
		テキスト部分に 主観を含まない	blog更新します 😊	7 (35.0)	
		テキスト部分に主観を 含まない広告ツイート	シリアルナンバー譲ります 😊 4枚ある	3 (15.0)	
		日本語ではない	시키프레 시키프레 😊	4 (20.0)	
	怒る😡	テキスト部分と 絵文字の主観が逆	可愛いすぎ 😡	25 (71.4)	
		テキスト部分に 主観を含まない	ぶんしゅかしよう 😡	10 (28.6)	
		🔴	テキスト部分と 絵文字の主観が逆	あいらびゅ 🔴	9 (45.0)
			テキスト部分に 主観を含まない	ですよな 🔴	10 (50.0)
			日本語ではない	para d atormentar o otacu fedido caveira fudida 🔴	1 (5.0)
	絵文字を含まない ツイート		肯定的極性	山下さんいいね	18 (81.8)
		否定的極性	涙腺が死ぬ	4 (18.2)

図1 CNNモデルのアーキテクチャ

等の文字列), URLs, および, “RT” の文字列を削除し, ツイート中のテキスト部分のみを処理対象とする。次に, ツイートテキストに対して形態素への分割を行う際には, 形態素解析システムとして MeCab を用い, 形態素解析用辞書として, ウェブ上の言語資源を用いて構築された mecab-ipadic-NEologd 辞書¹¹を用いる。また, 単語分散表現としては, 日本語 Wikipedia 中の単語および Wikipedia 記事となっているエンティティを対象とした分散表現ベクトルである「日本語 Wikipedia エンティティベクトル」^{12,13}を用いる。

また, 本論文においては, Kim [2] におけるモデルのうち, 比較の実装が容易な 1 つのチャンネルを用いる以下の 3 種類のモデルを採用し, その評価結果を示す。

- **CNN-rand**: 単語に対して分散表現の初期値を無作為に付与し, モデルの訓練に伴って単語の分散表現が更新される。
- **CNN-static**: 単語に対する分散表現の初期値として「日本語 Wikipedia エンティティベクトル」を付与する。「日本語 Wikipedia エンティティベクトル」においてエントリが

存在せず, 分散表現の初期値を無作為に与えられた単語も含めて, 全ての単語の分散表現を初期値に固定したままで, モデルの訓練が行われる。

- **CNN-non-static**: 単語の分散表現の初期値は CNN-static モデルと同じであるが, モデルの訓練に伴って単語の分散表現が更新される。

4. 評価事例

本節では, 本論文の評価実験において共通に用いる評価事例について述べる。

本論文の評価事例は, 2 節において収集した全 8,405,298 ツイート (表 1 の「総数」欄) を母集団とする。このうち, 2 節において分析対象としたツイートを除外したツイート集合から無作為に抽出したツイートに対して, 人手で肯定, 否定, 中立の三値極性を判定した。¹⁴ 人手による極性判定結果として, 肯定極性 1,000 ツイート, 否定極性 1,000 ツイート, 中立極性 1,000 ツイートを収集し, これらの極性判定済みツイートを 2 等分して, ツイート集合 $TrEv_{man}$, および, Ev とする。このうち, ツイート集合 $TrEv_{man}$ (肯定極性 500 ツイート, 否定極性 500 ツイート, 中立極性 500 ツイート) を, 次節の CNN モデル・SVM モデルの性能比較において訓練・評価事例として用いる。¹⁵ 一方, ツイート集合 Ev (肯定極性 500 ツイート, 否定極性 500 ツイート, 中立極性 500 ツイート) は, 次節の評価および 6 節の評価において, 共通の評価事例として用いる。¹⁶

表 1 において「絵文字を含まないツイート」7,597,377 ツイートを収集する際に除外された約 1,000 種類の絵文字を削除対象とした。

¹¹ <https://github.com/neologd/mecab-ipadic-neologd>

¹² http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/

¹³ 各語のベクトルの次元数は 200 であり, 分散表現を持たない語に対しては, 初期分散表現として, $-0.25 \sim 0.25$ の範囲から無作為に選ばれた数値を各次元値とするベクトルを用いる。

¹⁴ 本論文においては, 一人の作業によってツイートの極性判定作業を行ったが, そのうちの 1,000 ツイート分を対象として極性判定を二者で行った結果における Cohen の κ 係数 [6] の値は 0.785 でありほぼ一致という結果となった。

¹⁵ その他, 6 節の「絵文字の有無に基づき主観極性を自動同定したツイートをを用いたモデル訓練とその評価」において, 比較対象として用いる「手動で収集した訓練事例によって訓練した主観極性同定 CNN モデル M_{man} 」の epoch 数決定用開発事例として用いる。

¹⁶ その他, 次節の「主観極性を人手で同定したツイートをを用いたモデル訓練とその評価」において, CNN の epoch 数決定用, および, SVM のカーネル関数・ハイパーパラメータ調整用開発事例として用いる。

5. 主観極性を人手で同定したツイートを用いたモデル訓練とその評価

本節では、まず、人手で肯定・否定・中立極性を同定したツイート集合を収集し、これを訓練事例としてツイートに対する肯定・否定・中立の主観極性を同定するモデル M_{man} を訓練する方式およびその評価結果について述べる。

5.1 訓練・評価手順

本節では、主観極性判定モデルとして、CNN および SVM¹⁷ を用いて主観極性判定モデルの訓練・開発・評価を行う。ツイート集合として前節で述べた $TrEv_{man}$ を用いて 10 分割交差検定を行うことにより、モデルの訓練・開発・評価を行う。ただし、CNN において、ドロップアウト率、L2 正則化係数、損失最小となる epoch 数の決定用、および、SVM のカーネル関数・ハイパーパラメータ調整用開発事例として前節で述べたツイート集合 Ev を用いる。¹⁸

本論文では、肯定・否定・中立の各クラスの評価事例を同定する場合の再現率・適合率によって各モデルの性能を評価する。より具体的には、CNN の softmax 関数の確率値、および、SVM の信頼度の確率値に対して下限値を設定し、この下限値を変化させることによって、再現率・適合率曲線をプロットし、この曲線の上下関係に基づいて各モデルの性能を評価する。

なお、CNN においては、例えば、肯定クラスを同定するモデルを訓練する場合には、訓練時におけるクラス分類としては、1) 肯定・否定・中立の三値分類、2) 肯定クラスに対して、否定・中立を合わせたクラスを識別する二値分類、3) 肯定クラスと否定クラスのみを識別する二値分類、および、4) 肯定クラスと中立クラスのみを識別する二値分類、の 4 通りが考えられる。¹⁹ 否定クラスの同定、中立クラスの同定においても、同様にそれぞれ 4 通りとなる。さらに、3.2 節で述べた CNN-rand、CNN-static、および、CNN-non-static の 3 通りの分散表現のモデル化が考えられる。そこで、本論文では、訓練時クラス分類全 4 通り × 分散表現のモデル化 3 通りの合計 12 通りのモデル化に対して、本節の評価結果における性能が最もよいモデル化を採用する。

同様に、SVM においても、訓練時におけるクラス分類として全 4 通りのモデル化が考えられる。そこで、SVM においても、本節の評価結果における性能が最もよいモデル化を採用する。

17 SVM パッケージとしては、LIBSVM [7,8] を用いる。SVM 適用時には、3.2 節で述べた手順に従い、ツイート中のテキスト部分のみを処理対象とした後、形態素単位に分割し、各形態素を SVM の素性として用いる。

18 ドロップアウト率の値は 0.1 から 0.9 まで 0.1 刻みで、一方、L2 正則化係数の値は 0.5 から 5 まで 0.5 刻みで、それぞれ増加させた。

19 3) および 4) においては、二値分類から除外されたクラス（具体的には、3) の中立クラス、および、4) の否定クラス）について、訓練事例集合中ではそれらの除外クラスの事例は対象外として訓練には使用しない。一方、評価事例集合においては、除外クラスの事例を含め全ての事例を使うが、クラス同定において付与されるクラスは、除外クラスを除く二値（具体的には、3) では「肯定クラス」または「否定クラス」、4) では「肯定クラス」または「中立クラス」となる。

図2 主観極性を人手で同定した訓練事例を用いた場合の CNN・SVM の性能比較結果

そして、最後に、CNN モデルと SVM モデルの性能を比較し、より高性能なモデルを選定し、6 節の「絵文字の有無に基づき主観極性を自動同定したツイートを用いたモデル訓練とその評価」を行う。

5.2 評価結果

前節の手順に従い、各モデル化の評価を行った結果、

CNN においては、肯定・否定・中立のいずれの極性においても、前節の 1) 肯定・否定・中立の三値分類を行うモデルが最も高性能となった。ただし、分散表現のモデル化としては、肯定および否定極性においては CNN-rand（ドロップアウト率 0.8、L2 正則化係数 0.5）を用いた場合が、また、中立極性においては CNN-nonstatic（ドロップアウト率 0.9、L2 正則化係数 5）を用いた場合が、それぞれ最も高性能となった。また、SVM においても、肯定・否定・中立いずれの極性においても、前節の 1) 肯定・否定・中立の三値分類を行った場合が最も高性能となった。これらのモデルの評価結果をプロットした結果を図 2 に示す。

この結果から分かるように、CNN モデルと SVM モデルの性能比較においては、肯定クラス、否定クラス、および、中立クラスのいずれにおいても、CNN モデルの性能が SVM モデルの性能を上回った。具体的には、肯定極性、および、中立極性においては、再現率・適合率とも、CNN モデルが SVM モデルを上回っている。否定極性においては、再現率において SVM モデルが CNN モデルをやや上回るものの、適合率において CNN モデルが SVM モデルを大幅に上回っている。以上をふまえて、6 節においては、肯定クラス、否定クラス、中立クラスのいずれにおいても、本節で採用した CNN モデルを用いて「絵文字の有無に基づき主観極性を自動同定したツイートを用いたモデル訓練とその評価」を行う。

5.3 考察

図 2 の評価結果においては、否定極性の再現率において SVM が CNN をやや上回ったが、その差はわずかであり、それ以外の肯定・中立極性の再現率、および、肯定・否定・中立極性の適合率においては、CNN が SVM を上回っている。そこで、まず、CNN が再現率・適合率において SVM を上回る原因の一

端となる事例および現象について考察するため、以下の(例1)～(例3)をとりあげる。

(例1) いえいえ！最近ハマった新参なのでまだよくわかってないので、反応いただけて嬉しいです！よろしくお願いします。

(例2) 痛みはまだあるけれど腫れてもないので大丈夫だと思います！ありがとう！

(例3) 改善すればより良い物になる。それは間違い無からう。

(下線部は典型的な肯定・否定極性語)

(例1)では、CNNは「肯定極性」と正しく判定したのに対して、SVMは「否定極性」と誤判定した。図2の評価結果においては、(例1)と同様の事例が一定数存在し、結果的に肯定極性事例の正答数においてCNNがSVMを上回り、CNNにおいて肯定極性の再現率・適合率に寄与する要因の1つとなった。一方、人手付与極性が肯定の事例の誤判定に着目すると、否定極性へと誤判定する事例数においてはSVMがCNNを上回り、SVMにおいて否定極性の適合率を下げる要因の1つとなった。この事例では、典型的な肯定極性語²⁰（「よく」、「嬉しい」）、および、否定極性語（「いえいえ」、「ハマった」、「まだ」、「ない」）の両方を一文内に含んでおり、SVMは否定極性語の影響を強く受けて「否定極性」と誤判定したのに対して、CNNは単語埋め込みをふまえることにより「肯定極性」と正しく判定できたと考えられる。同様に、(例2)では、CNNは「肯定極性」と正しく判定したのに対して、SVMは「中立極性」と誤判定した。図2の評価結果においては、(例2)と同様の事例が一定数存在し、(例1)での議論をふまえると、CNNにおいて肯定極性の再現率・適合率に寄与するとともに、SVMにおいて中立極性の適合率を下げる要因の1つとなった。この事例でも、典型的な肯定極性語（「大丈夫」、「ありがとう」）、および、否定極性語（「痛み」、「腫れて」、「ない」）の両方を一文内に含んでおり、SVMは肯定・否定両極性語の影響を受けて「中立極性」と誤判定したのに対して、CNNは単語埋め込みをふまえることにより「肯定極性」と正しく判定できたと考えられる。また、(例3)では、CNNは「中立極性」と正しく判定したのに対して、SVMは「否定極性」と誤判定した。図2の評価結果においては、(例3)と同様の事例が一定数存在し、(例1)での議論をふまえると、CNNにおいて中立極性の再現率・適合率に寄与するとともに、SVMにおいて否定極性の適合率を下げる要因の1つとなった。この事例でも、典型的な肯定極性語（「改善」、「良い」）、および、否定極性語（「間違い」、「無からう」）の両方を一文内に含んでいるが、SVMは否定両極性語の影響を受けて「否定極性」と誤判定したのに対して、CNNは肯定・否定両極性語の単語埋め込みをふまえることにより「中立極性」と正しく判定できたと考えられる。なお、以上の(例1)～(例3)と同様の事例は一定数存在するが、それ

ら以外にも、人手で付与された極性が肯定・否定・中立であるか否かに関わらず、典型的な肯定・否定極性語を文中に含む事例も含まない事例も、それぞれ一定数存在する。これらの事例に対するCNN・SVMによる判定の正誤の数を総合した結果、肯定・中立極性の再現率、および、肯定・否定・中立極性の適合率において、CNNがSVMを上回る結果となった。

次に、総合的に性能の高いCNNに着目して、肯定・否定・中立の各極性間で性能の比較をすると、図2においては、再現率・適合率とも、肯定・否定極性の間ではほぼ同等の性能となっているが、中立極性においては、再現率・適合率とも肯定・否定極性を下回っている。このことを考察するために、以下では、訓練・評価事例集合 *TrEv_{man}* 中において、肯定・否定・中立の各極性間の誤判定数に着目する。まず、再現率について、図2のプロット右端において、人手付与極性が肯定、否定、中立の事例における誤判定数は、それぞれ、約110前後、約130前後、約200前後であった。ここで、この誤判定数が中立極性において突出して多くなる理由は、肯定・否定極性間の誤判定数が約40弱と少ないのに対して、肯定・中立極性間、および、否定・中立極性間の誤判定数が約70～90前後と倍前後であるからである。適合率についても、図2のプロット右端における各極性への誤判定数は、肯定約120前後、否定約130前後に対して、中立約180前後である。ここでも、上述の再現率における議論と同様に、肯定・否定極性間の誤判定数が約40弱と少ないのに対して、肯定・中立極性間、および、否定・中立極性間の誤判定数が約90弱と倍以上であるため、中立極性への誤判定数が突出して多くなっている。これらの原因として、肯定・否定極性の文においては、極性判定の明示的な手がかりとなる肯定・否定極性語が多く含まれる傾向にあるのに対して、中立極性の文においては、それらの極性語が含まれていなかったり、逆に、含まれていても、極性判定の誤判定の原因となる場合が少なからずあることが挙げられる。以下の(例4)～(例9)では、肯定・否定極性語が誤判定の原因となる例を中心に、肯定・否定・中立極性間でCNNが誤判定をした事例を示す。

(例4) もう残念だったけど可愛すぎるぞ！

(肯定極性語「可愛」にも関わらず、否定極性語「残念」、「すぎる」が原因となって、肯定極性ツイートを「否定極性」とであると誤判定)

(例5) 美味しいよねー！食べすぎるとオエツとくるけど。

(肯定極性語「美味しい」にも関わらず、否定極性語「すぎる」、「オエツ」が原因となって、肯定極性ツイートを「中立極性」とであると誤判定)

(例6) 見て来たw怖かったあwww

(否定極性語「怖かった」にも関わらず、否定極性ツイートを「肯定極性」とであると誤判定)

(例7) 大好きだった物すら食べれない。大好きだった遊びすらできない。僕のやる気はどこに行ったのかな。

(否定極性語「ない」にも関わらず、肯定極性語「大好き」、「やる気」が原因となって、否定極性ツイートを「中立極性」とであると誤判定)

(例8) 2割でいいからくれよ。

20 本節の考察においては、典型的な肯定極性語あるいは否定極性語であると想定される語を例示している。しかし、これらの極性語はいずれも、各語が持つ一般的な印象に基づいて例示したものであり、本論文の評価実験においてその極性を数値化した結果に基づくものではない。

(肯定極性語「いい」が原因となって、中立極性ツイートを「肯定極性」とであると誤判定)

(例9) カステラじゃないの...?

(否定極性語「ない」が原因となって、中立極性ツイートを「否定極性」とであると誤判定)

(下線部は典型的な肯定・否定極性語)

これらの事例は、CNN が誤判定を起こす典型例の一例ではあるが、これら以外にも、否定極性語が含まれないにも関わらず「否定極性」とであると誤判定する例、逆に、肯定極性語が含まれないにも関わらず「肯定極性」とであると誤判定する例、肯定・否定極性語が含まれないにも関わらず「中立極性」とであると正しく判定できない例等も一定数存在する。

6. 絵文字の有無に基づき主観極性を自動同定したツイートを用いたモデル訓練とその評価

本節では、絵文字の有無に基づき主観極性を自動同定したツイートをを用いて訓練した主観極性同定モデル M_{auto} の性能評価を行い、5 節で述べた、人手で主観極性を同定したツイートをを用いて訓練した主観極性同定モデル M_{man} との間で性能の比較評価を行った結果について述べる。本節の性能評価においては、5 節の性能評価結果をふまえて、肯定・否定・中立の三値分類を行うモデルを用いることとする。ただし、分散表現のモデル化としては、肯定および否定極性同定においては CNN-rand (ドロップアウト率 0.8, L2 正則化係数 0.5) を、一方、中立極性同定においては CNN-nonstatic (ドロップアウト率 0.9, L2 正則化係数 5) を、それぞれ用いる。なお、これらのモデルにおいては、計算時間短縮のため、5 節「主観極性を人手で同定したツイートをを用いたモデル訓練とその評価」において開発事例を用いて決定したドロップアウト率、および、L2 正則化係数の値をそのまま用いる。

6.1 訓練・開発事例の収集

本節では、絵文字の有無に基づき主観極性を自動同定したツイートを収集し、主観極性同定モデルの訓練・開発事例として用いる。具体的には、表 2 の「3 種類の絵文字のうち、1 種類のみを含み、他の 2 種類の絵文字を含まないツイート」の項目において、絵文字 のみを含む 387,412 ツイートから無作為に抽出した 45,404 ツイートを、肯定性クラスのツイートとする。一方、表 2 の「3 種類の絵文字のうち、1 種類のみを含み、他の 2 種類の絵文字を含まないツイート」の項目において、絵文字 のみを含む 29,765 ツイート、および、絵文字 のみを含む 15,640 ツイートの和集合から無作為に抽出した 45,404 ツイートを、否定極性クラスのツイートとする。さらに、表 2 の「絵文字を含まないツイート」の項目の 7,597,377 ツイートから無作為に抽出した 45,404 ツイートを、中立極性クラスのツイートとする。²¹

21 本論文の実験設定においては、肯定・否定・中立の各極性間で均等な事例数のもとで訓練・開発事例を構築する。一方、本論文で収集

6.2 訓練・評価手順

本節では、「主観極性を自動同定したツイートをを用いて訓練したモデル M_{auto} 」、および、「主観極性を人手で同定したツイートをを用いて訓練したモデル M_{man} 」の訓練・開発・評価を行う手順の詳細について述べる。これらのモデルの評価においては、4 節で述べたツイート集合 E_v を共通に用いる。特に、モデル M_{man} の訓練・開発・評価においては、ツイート集合 E_v を用いた 10 分割交差検定を行うとともに、10 分割交差検定の全 10 回の訓練のうちの 1 回ごとに、4 節で述べたツイート集合 $TrEv_{man}$ を開発事例集合として用いる。一方、モデル M_{auto} の訓練・開発・評価においては、前節で自動収集したツイート集合をなるべく多く用いて性能の高いモデルを訓練するために、互いに重複のない訓練事例・開発事例の選定を無作為に行い、モデルの訓練と epoch 数の決定をした後、4 節で述べたツイート集合 E_v を用いたモデルの評価を行う過程を 10 回繰返し、その 10 回のマクロ平均を求める。

6.2.1 主観極性を自動同定したツイートをを用いて訓練したモデル M_{auto}

前節で自動収集した肯定・否定・中立極性各 45,404 ツイートに対して、各極性ごとの訓練事例を無作為に N_{auto} ($\leq 22,702$) ツイート選定し、モデル M_{auto} の訓練事例として用いる。さらに、各極性ごとの母集団 45,404 ツイートにおいて、訓練事例の N_{auto} ツイートを除外した $45,404 - N_{auto}$ ツイート中から、 N_{auto} ($\leq 22,702$) ツイートを無作為に選定し、損失最小となる epoch 数の決定用開発事例として用いる。 $N_{auto} = 22,702, 3,000, 1,500$ の 3 通り²² の各々に対応するモデルを「**自動**」訓練 + 開発 = 68,106 + 68,106, 「**自動**」訓練 + 開発 = 9,000 + 9,000, および, 「**自動**」訓練 + 開発 = 4,500 + 4,500」とし、上述の無作為選定、モデル訓練、 E_v による評価を 10 回繰返し、その 10 回のマクロ平均を求め、3 通りのモデル M_{auto} の評価を行う。

6.2.2 主観極性を人手で同定したツイートをを用いて訓練したモデル M_{man}

比較対象のモデル M_{man} の訓練・開発・評価においては、ツイート集合 E_v (肯定極性 500 ツイート, 否定極性 500 ツイート, 中立極性 500 ツイート) を用いた 10 分割交差検定を行う。ただし、交差検定における 10 回のモデル訓練の各 1 回においては、各極性ごとの訓練事例を無作為に N_{man} (≤ 450) ツイート選定し、モデル訓練を行う。その際、ツイート集合 $TrEv_{man}$ (肯定極性 500 ツイート, 否定極性 500 ツイート, 中立極性 500 ツイート) から、各極性ごとに N_{man} (≤ 450) ツイートを

した事例集合においては、表 2 に示すように、「怒る」に対応する否定極性が最も事例数が少なく、総数 45,405 個となる。したがって、本論文においては、「肯定・否定・中立極性事例数として、それぞれ 45,404 個」というのが、利用可能な最大規模の事例数となる。

22 $N_{auto} = 300, 150$ の 2 通りの各々に対しても同様にモデル M_{auto} の評価を行ったが、モデル M_{man} と比較して相対的に適合率が劣る結果となったため、本論文では、 $N_{auto} = 22,702, 3,000, 1,500$ の 3 通りに対する評価結果のみを示す。

選定し²³, 損失最小となる epoch 数の決定用開発事例として用いる. $N_{man} = 450, 225, 90, 45$ の 4 通り²⁴ の各々に対応するモデルを「[手動] 訓練 + 開発 = 1,350 + 1,350」, 「[手動] 訓練 + 開発 = 675 + 675」, 「[手動] 訓練 + 開発 = 270 + 270」, および, 「[手動] 訓練 + 開発 = 135 + 135」として上述の手順を行い, 4 通りのモデル M_{man} の評価を行う.

6.3 評価結果

4 通りのモデル M_{man} の評価結果を, それぞれ, 図 3 の「訓練 + 開発 = 1,350 + 1,350」, 「訓練 + 開発 = 675 + 675」, 「訓練 + 開発 = 270 + 270」, 「訓練 + 開発 = 135 + 135」に示す (簡略化のため, 図 3 中の表記においては, [手動] を省略する). この結果から分かるように, 肯定極性クラス (図 3(a)), 否定極性クラス (図 3(b)), 中立極性クラス (図 3(c)) のいずれにおいても, 全体として, 訓練・開発に用いる事例数が増えるとともに, 性能の改善が見られた. 次に, 図 4 においては, 3 通りのモデル M_{auto} の評価結果を, それぞれ, 「[自動] 訓練 + 開発 = 68,106 + 68,106」, 「[自動] 訓練 + 開発 = 9,000 + 9,000」, および, 「[自動] 訓練 + 開発 = 4,500 + 4,500」に示し, これを, 「[手動] 訓練 + 開発 = 270 + 270」, および, 「[手動] 訓練 + 開発 = 135 + 135」の 2 通りのモデル M_{man} の評価結果と比較する. この結果から分かるように, 肯定極性クラス (図 4(a)), 否定極性クラス (図 4(b)), および, 中立極性クラス (図 4(c)) のいずれにおいても, 「[自動] 訓練 + 開発 = 68,106 + 68,106」, 「[自動] 訓練 + 開発 = 9,000 + 9,000」, および, 「[自動] 訓練 + 開発 = 4,500 + 4,500」のモデル M_{auto} の性能が, おおよそ, 「[手動] 訓練 + 開発 = 270 + 270」と「[手動] 訓練 + 開発 = 135 + 135」の 2 つのモデル M_{man} の性能の間におさえられている, あるいは, 2 つのモデル M_{man} の性能とほぼ同等である. このことから, 本論文の方法により絵文字を含むツイート, および, 絵文字を含まないツイート, 合計 9,000~136,212 ツイートを自動収集したものを訓練・開発事例として訓練した主観極性同定モデルの性能は, 主観極性を人手で同定した訓練・開発事例 270~540 ツイートを用いて訓練した主観極性同定モデルの性能に匹敵することが分かる.²⁵ 以上の評価結果をふまえると, 本論文の方法により絵文字を含むツイート, および, 絵文字を含ま

ないツイートを自動収集して訓練・開発事例として用いるアプローチが有効であることが分かる.

7. 関連研究

本節では, ツイートを対象として主観同定を行う手法についての関連研究について, いくつかの観点のもとで, 本論文における提案手法との比較を行う.

以下では, まず, モデル訓練用の訓練事例の収集において, 手動で収集する方式を採用する研究事例, および, 何らかの手がかりに基づき自動で収集する方式を採用する研究事例について, それぞれ述べる.

ツイートの主観同定タスクにおいて, モデル訓練用の訓練事例を手動で収集する取り組みの代表例として, SemEval の Sentiment Analysis in Twitter タスクにおいて配布された訓練事例およびそれを利用した研究事例 [9–14] が挙げられる. SemEval の Sentiment Analysis in Twitter タスクにおいては, 人手で主観極性が同定された訓練用ツイート集合を作成し, タスク参加者に対して配布している. それ以外でも, SemEval の Sentiment Analysis in Twitter タスクにおいて配布されたデータセットとは別に, 各研究事例において個別に訓練・評価用ツイート集合を収集し, 人手で主観極性を同定したデータセットもいくつか存在しており, それらの個別のデータセットを利用する事例 [15–18] も知られている.

一方, 文献 [1, 19–21] においては, ハッシュタグ, あるいは, 記号 “:”²⁶, および, “:(”²⁷ 等の何らかの手がかりを利用することにより, ツイートの主観同定モデル訓練事例を自動収集するアプローチを提案している. これらのアプローチでは, 例えば, 文献 [20] においては, ハッシュタグに対して事前に肯定・否定・中立の極性を規定しておき, それらの極性ごとに, ハッシュタグを含むツイートを自動収集して, ツイートの主観同定モデル訓練事例としている. 同様に, 文献 [1, 19–21] においては, 記号 “:”²⁸, および, “:(”²⁹ を手がかりとして, ツイートの主観同定モデル訓練事例を自動収集している.

これに対して, 本論文では, 主観極性を持つ高頻度な絵文字を手がかりとしてツイートの主観同定モデル訓練事例を自動収集するアプローチを提案している. 具体的には, 本論文では, 絵文字 を含むツイートのテキスト部分が肯定極性を持ち, 絵文字 または を含むツイートのテキスト部分が否定極性を持ち, 絵文字を含まないツイートのテキスト部分が中立極性を持つと仮定して, 訓練事例の自動収集を行う点が異なる. また, 本論文では, 自動収集された訓練事例の性能を見積もることを目的として, 手動で主観極性を同定した訓練事例によって訓練された主観極性同定モデルを比較対象として, 自動収集された訓練事例によって訓練された主観極性同定モデルの評価を行っており, この点において本論文の新規性がある.

一方, 絵文字および顔文字を含むテキストについての関連研究の観点からは, 文献 [22] においては, 4 種類の主観情報の教師ありクラス分類タスクにおいて, 深層学習を用いた教師なし素性学習用の訓練事例として, 絵文字を含むマイクロブログ投稿を利用しその有効性を報告している. 文献 [17] においては, 約 1,600 種類の絵文字に対する約 6,000 個の説明文, および,

23 10 分割交差検定と同様の手順により, 各極性 500 ツイートを 10 分割して, 各事例が選ばれる回数が均等になるように開発事例集合を作成する.

24 $N_{man} = 405, 360, 315, 270, 180, 135$ の 6 通りの各々に対しても同様にモデル M_{man} の評価を行ったが, モデル M_{auto} と比較して相対的に適合率が大幅に上回る結果となった. そこで, 本論文では, モデル M_{auto} と同等の適合率となるモデル M_{man} として, $N_{man} = 90, 45$ の 2 通りに対する評価結果を示し, モデル M_{auto} の適合率を大幅に上回るモデル M_{man} としては, $N_{man} = 450, 225$ の 2 通りに対する評価結果のみを示す.

25 ただし, 図 4 の評価結果から分かるように, 「[自動] 訓練 + 開発 = 68,106 + 68,106」, 「[自動] 訓練 + 開発 = 9,000 + 9,000」, および, 「[自動] 訓練 + 開発 = 4,500 + 4,500」の 3 つのモデル M_{auto} の間に大きな性能の差はない. このことから, 本論文の問題設定において自動収集されたツイートを訓練・開発事例として主観極性同定モデルを訓練するアプローチにおいては, 訓練・開発事例の数がある一定数以上になると, 主観極性同定モデルの性能がそれ以上大幅には改善しない可能性があることが分かる.

(a) 肯定極性クラス

(b) 否定極性クラス

(c) 中立極性クラス

図3 主観極性を人手で同定した訓練事例を用いた場合のモデル評価結果

Google News を用いて訓練した分散表現を用いることにより、絵文字に対する分散表現 emoji2vec を訓練する手法を提案している。文献 [15] においては、ツイートに対して、絵文字以外の周辺語情報を素性とする教師あり学習によって、特定の国に対する主観を同定する手法を提案している。この手法においては、ツイート中の主観情報の対象が、料理等、国以外のものとなる場合を除外する手法を適用した後、特定の国に対する主観を同定している。また、文献 [21] においては、顔文字を含むツイートを訓練データとして LSTM モデルを利用し、主観の極性を分類する手法を提案している。

また、ツイートの主観極性についての関連研究の観点からは、文献 [16] においては、主観の対象となる語と主観語との間の構文的関係を考慮して、主観の対象となる語に応じて適応的

(a) 肯定極性クラス

(b) 否定極性クラス

(c) 中立極性クラス

図4 絵文字の有無に基づき主観極性を自動同定した訓練事例を用いた場合のモデル評価結果

に作用する RNN によって、ツイートの主観を同定する手法を提案している。文献 [14] においては、トピックモデルによって話題に分割されたツイート集合に対して SVM による主観分類器を適用する手法を提案している。また、文献 [18] においては、ツイートに含まれる複数の対象に対して、対象別に主観を同定するタスクに取り組んでいる。

8. おわりに

本論文では、ツイート中の主観極性同定モデルの訓練事例の作成にあたって、絵文字の有無および絵文字の種類を手がかりとして、人手を介さず自動で訓練用ツイートを収集し、これを用いてツイート中の主観極性同定モデルを訓練する方式を提案

し、その有効性を示した。また、ツイート中の主観極性同定モデルとしては、CNN および SVM の比較を行った。その結果、CNN による主観極性分類性能の方が SVM よりも優れていることを示した。また、絵文字の有無に基づき主観極性を自動同定したツイートを用いて訓練した主観極性同定モデル M_{auto} に対して、人手で主観極性を同定したツイートを用いて訓練した主観極性同定モデル M_{man} との間で性能の比較評価を行った結果を示した。その結果においては、絵文字を含むツイート、および、絵文字を含まないツイート、合計 9,000~136,212 ツイートを自動収集したものを訓練・開発事例として訓練した主観極性同定モデルの性能が、主観極性を人手で同定した訓練・開発事例 270~540 ツイートを用いて訓練した主観極性同定モデルの性能に匹敵することが分かった。この評価結果をふまえて、本論文では、絵文字を含むツイート、および、絵文字を含まないツイートを自動収集して訓練・開発事例として用いるアプローチが有効であることを示した。今後の課題として、BERT [23] 等、大規模言語資源を用いて事前訓練された分類モデルを適用し、本論文で述べた CNN との間で性能を比較することが挙げられる。

参 考 文 献

- [1] N. Kalchbrenner, E. Grefenstette, and P. Blunsom: "A convolutional neural network for modelling sentences," *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, pp. 655-665, 2014.
- [2] Y. Kim: "Convolutional neural networks for sentence classification," *Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, pp. 1746-1751, 2014.
- [3] P. Wang, J. Xu, B. Xu, C. Liu, H. Zhang, F. Wang, and H. Hao: "Semantic clustering and convolutional neural network for short text categorization," *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int. Joint Conf. on Natural Language Processing*, Beijing, pp. 352-357, 2015.
- [4] D. P. Kingma and J. Ba: "Adam: A method for stochastic optimization," *Proc. of the 3rd Int. Conf. on Learning Representations (ICLR)*, May 7-9, San Diego, 2015.
- [5] M. D. Zeiler: "ADADELTA: an adaptive learning rate method," Vol. abs/1212.5701, arXiv preprint, arXiv:1212.5701v1, 2012.
- [6] J. Cohen: "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 37-46, 1960.
- [7] C.-C. Chang and C.-J. Lin: "LIBSVM: A library for support vector machines," *ACM Trans. on Intelligent Systems and Technology*, Vol. 2, Article No. 27, pp. 27:1-27:27, 2011.
- [8] C.-C. Chang and C.-J. Lin: LIBSVM, <http://www.csie.ntu.edu.tw/~cjlin/libsvm> [accessed Dec. 1, 2019]
- [9] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov: "SemEval-2016 Task 4: Sentiment Analysis in Twitter," *Proc. of the 10th Int. Workshop on Semantic Evaluation (SemEval-2016)*, Jun. 16-17, San Diego, pp. 1-18, 2016.
- [10] P. Nakov, S. Rosenthal, Z. Kozareva, V. Stoyanov, A. Ritter, and T. Wilson: "SemEval-2013 Task 2: Sentiment Analysis in Twitter," *Proc. of the 7th Int. Workshop on Semantic Evaluation (SemEval-2013)*, Jun. 14-15, Atlanta, pp. 312-320, 2013.
- [11] S. Rosenthal, N. Farra, and P. Nakov: "SemEval-2017 Task 4: Sentiment Analysis in Twitter," *Proc. of the 11th Int. Workshop on Semantic Evaluation (SemEval-2017)*, Aug. 3-4, Vancouver, pp. 502-518, 2017.
- [12] S. Rosenthal, P. Nakov, S. Kiritchenko, S. Mohammad, A. Ritter, and V. Stoyanov: "SemEval-2015 Task 10: Sentiment Analysis in Twitter," *Proc. of the 9th Int. Workshop on Semantic Evaluation (SemEval-2015)*, Jun. 4-5, Denver, pp. 451-463, 2015.
- [13] S. Rosenthal, A. Ritter, P. Nakov, and V. Stoyanov: "SemEval-2014 Task 9: Sentiment Analysis in Twitter," *Proc. of the 8th Int. Workshop on Semantic Evaluation (SemEval-2014)*, Aug. 23-24, Dublin, pp. 73-80, 2014.
- [14] B. Xiang and L. Zhou: "Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training," *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, Jun. 23-25, Baltimore, pp. 434-439, 2014.
- [15] N. Chambers, V. Bowen, E. Genco, X. Tian, E. Young, G. Harihara, and E. Yang: "Identifying political sentiment between nation states with social media," *Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing*, Lisbon, pp. 65-75, 2015.
- [16] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu: "Adaptive recursive neural network for target-dependent twitter sentiment classification," *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, Jun. 23-25, Baltimore, pp. 49-54, 2014.
- [17] B. Eisner, T. Rocktäschel, I. Augenstein, M. Bošnjak, and S. Riedel: "emoji2vec: Learning emoji representations from their description," *Proc. of the 4th Int. Workshop on Natural Language Processing for Social Media*, Nov. 1, Austin, pp. 48-54, 2016.
- [18] B. Wang, M. Liakata, A. Zubiaga, and R. Procter: "TDParse: Multi-target-specific sentiment recognition on twitter," *Proc. of the 15th Conf. of the European Chapter of the Association for Computational Linguistics*, Vol. 1, Apr. 3-7, Valencia, pp. 483-493, 2017.
- [19] A. Go, R. Bhayani, and L. Huang: "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, Vol. 1, No. 12, p. 1-6, 2009.
- [20] E. Kouloumpis, T. Wilson, and J. Moore: "Twitter sentiment analysis: The good the bad and the OMG!," *Proc. of the 5th Int. Conf. on Weblogs and Social Media*, Jul. 17-21, Catalonia, pp. 538-541, 2011.
- [21] X. Wang, Y. Liu, C. Sun, B. Wang, and X. Wang: "Predicting polarities of tweets by composing word embeddings with long short-term memory," *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int. Joint Conf. on Natural Language Processing*, Jul. 26-31, Beijing, pp. 1343-1353, 2015.
- [22] D. Tang, B. Qin, T. Liu, and Z. Li: "Learning sentence representation for emotion classification on microblogs," *Proc. 2nd CCF Int. Conf. on Natural Language Processing and Chinese Computing*, Nov. 15-19, Chongqing, pp. 212-223, 2013.
- [23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova: "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jun. 2-7, Minneapolis, pp. 4171-4186, 2019.

(2019 年 12 月 11 日 受付)
(2020 年 6 月 21 日 採録)

[問い合わせ先]

〒305-8573 茨城県つくば市天王台 1-1-1
筑波大学システム情報系
宇津呂 武仁
TEL: 029-853-6537/5427
E-mail: utsuro@iit.tsukuba.ac.jp

— 著 者 紹 介 —

りん ゆうちょう
林 友超 [非会員]

2017 年筑波大学大学院システム情報工学研究科博士前期課程修了。現在、同大学大学院システム情報工学研究科博士後期課程在学中。自然言語処理の研究に従事。

さい こうよく
崔 鴻翌 [非会員]

2020 年筑波大学大学院システム情報工学研究科博士前期課程修了。

うつろ たけひと
宇津呂 武仁 [正会員]

1994 年 京都大学大学院工学研究科博士課程修了。博士（工学）。1994 年 奈良先端科学技術大学院大学情報科学研究科助手、2000 年 豊橋技術科学大学情報工学系講師、2003 年 京都大学情報学研究科講師、2006 年 筑波大学大学院システム情報工学研究科助教授、准教授を経て 2012 年 筑波大学システム情報系教授、現在に至る。自然言語処理、ウェブマイニングの研究に従事。

Utilizing Emoji in Collecting Training Instances of a Model for Sentiment Analysis of Tweets

by

Youchao LIN, Hongyi CUI, and Takehito UTSURO

Abstract:

This paper proposes a method of automatically collecting a dataset for training a tweet sentiment analysis model. The proposed method automatically collects the training dataset based on the existence and the types of emoji. In this paper, we make a comparison between the sentiment analysis model trained with the automatically collected training dataset and the model trained with a training dataset that has manually identified sentiment labels. The evaluation result demonstrates that, in terms of the performance of the trained tweet sentiment analysis model, automatically collected 9,000–136,212 tweets correspond to 270–540 tweets with manually identified sentiment labels.

Keywords: Twitter, emoji, sentiment analysis

Contact Address: **Takehito UTSURO**

Faculty of Engineering, Information and Systems, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan
TEL: +81-29-853-6537/5427
E-mail: utsuro@iit.tsukuba.ac.jp