

単語のポジティブ・ネガティブでの出現比率に着目したツイートの極性分類

姫野晋之介 青野雅樹
 豊橋技術科学大学 情報・知能工学課程
 himeno@kde.cs.tut.ac.jp, aono@cs.tut.ac.jp

1. はじめに

Twitter や Facebook に代表されるマイクロブログは、最新のニュースやイベントを知るだけでなく、友人関係や社会的関係を形成するための大きな役割を担っている。全世界において、日々多くのユーザによる多種多様な書き込みがなされている。そのような中で、喜びや怒りに代表される感情表現が含まれる書き込みも大量に存在する。とくに Twitter は、全世界にアクティブユーザが 3 億人以上いることが知られており¹、その書き込みの容易さからリアルタイム性が強く、ユーザの意見が反映されやすいと考える。そのため、Twitter の感情を分析することはマーケティングや政治・経済などの分野に大きな影響を与えると考えられる。加えて、個々のユーザについても、新商品への感想やニュースへの意見など大規模な意見を必要とする場面での利便性があると考えられる。しかしながら、Twitter への書き込みは本文が 140 文字以内と短く、口語や俗語、略語が多く含まれるため、極性について分類するのは容易ではない。そのため、これらの性質、すなわち、大規模性、リアルタイム性、短文性、口語・略語性などの性質を考慮した高精度でツイートの極性を分類する手法が必要となる。

Twitter に関する感情分析の研究はこれまで様々な手法でなされてきた。多くは 3 極性(Positive, Neutral, Negative)への分類問題として扱われてきた。本研究では、同様の 3 極性分類問題として扱うが、大量の Twitter 本文の観察に基づき、ポジティブ・ネガティブな表現に出現する単語の出現比率に着目する手法を提案し、その実験結果を報告する。

2. 関連研究

Twitter に関する感情分析は情報検索の分野において注目されている。代表的な感情分析には以下の論文がある[1,12,13]。Twitter に投稿された書き込みはツイートと呼ばれる。先行研究の多くがツイートを Positive, Neutral, Negative の 3 極性に分類することを目的としている。ツイートの極性分類については Go らによって機械学習を用いた教師あり学習による手法が提案された[1]。Go らは Uni-Gram, Bi-Gram, 品詞情報に関する素性を提案した。これらの素性を基に

Naive Bayes, Maximum Entropy, Support Vector Machine の 3 種類の分類器を用いて極性分類実験を行った。近年では、Twitter の感情分析の国際コンテストのひとつである SemEval2016[10]にて、単語分散表現と文法的素性を組み合わせた手法[2]や、深層学習[3,4]を用いた手法が提案されている。また、教師なし学習による手法も提案[5,6]されている。感情分析の応用事例としては、政治選挙の動向[7]、自然災害の管理[8]、株式市場の変動予測[9]などがあげられる。

我々の手法は Go らと同様、Positive, Neutral, Negative の 3 極性のラベルが訓練データに付与された教師あり学習手法であり、以降の節で詳述する。

3. 提案手法

以下では我々が着目した極性に応じた単語の出現頻度に基づく手法に関して、3.1 節でポジティブ・ネガティブの出現比率を、3.2 節では、ポジティブ性とネガティブ性を導入する。3.3 節で素性の提案を、3.4 節で提案システムの流れを示す。

3.1 単語の出現比率

単語にはポジティブツイートでの出現比率が高い単語、ネガティブツイートでの出現比率が高い単語が存在すると考えられる。極性がラベリングされた過去の英語ツイートデータを用いて全単語の極性ごとの出現比率を算出する。全単語に対し、その単語が出現するポジティブツイート数、ネガティブツイート数を求める。それらを全ポジティブツイート数、全ネガティブツイート数で割ったものをその単語の出現比率とする。以下で T とは単語($Term$)を表す。

ポジティブツイートでの出現比率(T)

$$= \frac{T \text{ のポジティブツイートでの出現数}}{\text{全ポジティブツイート数}}$$

ネガティブツイートでの出現比率(T)

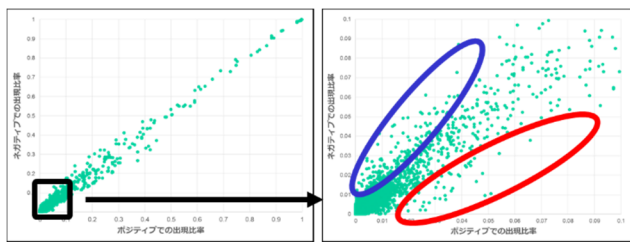
$$= \frac{T \text{ のネガティブツイートでの出現数}}{\text{全ネガティブツイート数}}$$

3.2 単語のポジティブ性・ネガティブ性

過去の英語ツイートデータ (表 2) において出現した全単語について、ポジティブツイートでの出現比率を X 軸、ネガティブツイートでの出現比率を Y 軸と

¹ <https://about.twitter.com/company>

してプロットしたグラフを図 1(a)に示す. 図 1(a)ではポジティブツイートでの出現比率とネガティブツイートでの出現比率に偏りのある単語は顕著ではない. しかし, 図 1(b)ではツイート全体で高い出現比率を持つ単語が多くプロットされている. 極性を持つような単語はツイート全体での出現比率は低いと考えられる. そのため X 軸, Y 軸の[0,0.1]の範囲を拡大したグラフを図 1(b)に示す.



(a) P-N グラフ (b) P-N グラフ (拡大)

図 1 英語ツイート(表 2)から得られた単語出現比率

図 1(b)では, 赤丸付近でポジティブツイートにて出現しやすい単語, 青丸付近でネガティブツイートにて出現しやすい単語が多く観測できる. この出現比率の偏りの程度をそれぞれポジティブ性・ネガティブ性と定義し, 以下の式(1)と式(2)でそれぞれ表現する.

ポジティブ性(T)

$$= \frac{\text{ポジティブツイートでの出現比率}(T)}{\text{ネガティブツイートでの出現比率}(T)} \quad (1)$$

ネガティブ性(T)

$$= \frac{\text{ネガティブツイートでの出現比率}(T)}{\text{ポジティブツイートでの出現比率}(T)} \quad (2)$$

これらの式を算出することで, 過去のツイートデータの全単語をそれぞれの極性での出現しやすさについてランキングすることが可能となる.

さらに, 単語だけでなく Bi-Gram, 単語ペアについても着目した. ここで Bi-Gram は文章中に出現した連続する 2 つの単語の組み合わせを指し, 単語ペアは出現順序に関係なく文章中に出現した 2 つの単語の組み合わせを指す. Bi-Gram, 単語ペアについても, 式(1)からポジティブ性, 式(2)からネガティブ性を算出し単語をそれぞれについてランキングする.

3.3 提案素性

式(1)と(2)に従い, ランキングした単語を用いた訓練データ・テストデータの素性ベクトル化について述べる. 各ツイートに対し, 後述する 3.3.1 で得られる素性ベクトルと, 3.3.2 から得られる素性ベクトルを連結したものを提案素性ベクトルとする.

3.3.1 出現頻度カウント

Bi-Gram, 単語ペアは出現頻度カウントによって素

性ベクトル化する. 各ツイートに対し, 2次元のベクトルを 1 つ考える. ツイート内にポジティブ性の上位 N 件の Bi-Gram, 単語ペアが含まれていれば, ポジティブに対応する要素の値をカウントアップする. 同様に, ネガティブ性の上位 N 件の Bi-Gram, 単語ペアが含まれていれば, ネガティブに対応する要素の値をカウントアップする. 含まれていない場合の要素の値は 0 とする. 出現頻度カウントの次元数は 2 である.

3.3.2 Bag-of-Words

単語は Bag-of-Words によって素性ベクトル化する. 各ツイートに対し, ポジティブ性, ネガティブ性の上位 N 件の単語に対応するベクトルを考える. ツイート内にポジティブ性, ネガティブ性の上位 N 件の単語が含まれていれば, 対応する要素の値をそのツイート内の出現回数とする. 含まれていない単語については対応する要素の値を 0 とする. 次元数は $2N$ である.

3.4 訓練システム・テストシステム

提案素性を用いて訓練システム・テストシステムを構築した. その流れを図 2 (a), (b)に示す.

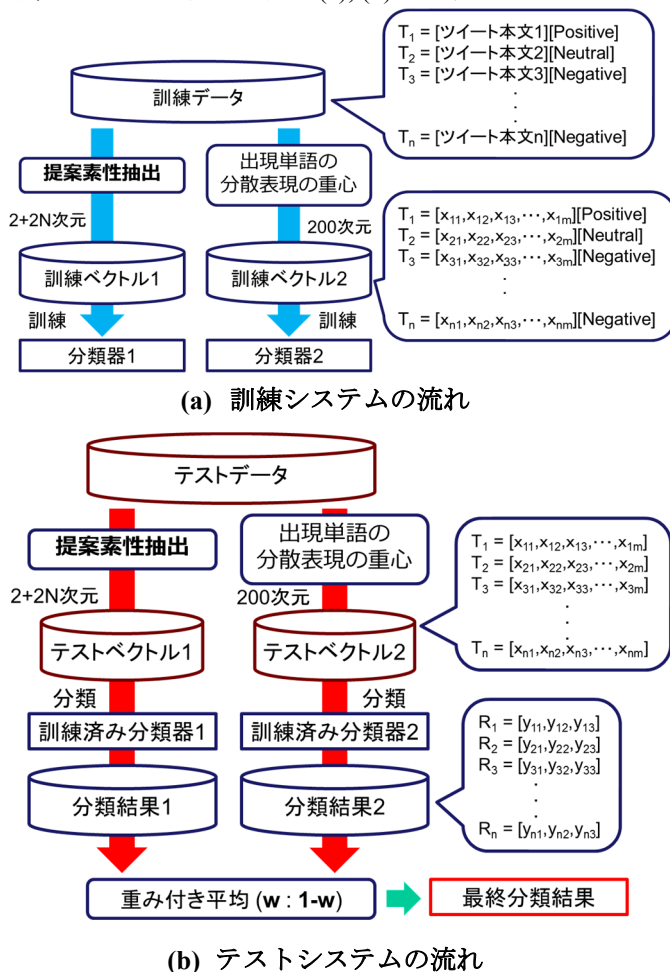


図 2 提案システムの流れ

図 2 (a)では, 訓練データの各ツイートを提案素性

によって素性ベクトル化する。その素性ベクトルを分類器1で訓練する。一方、単語分散表現である GloVe へも変換し、その素性ベクトルを分類器2で訓練する。テストデータの各ツイートも同様に素性ベクトル化し、それぞれを訓練済み分類器で分類する。本研究では分類器として LIBLINEAR²を用いている。それぞれの分類結果は3極性に対応する実数値が出力される。それぞれの分類器の出力を w と $(1-w)$ で重みを付け平均し、最大となる極性を最終分類結果とする。

4. 評価実験

4.1 データセット

本研究では、SemEval2016にて用いられた Train データ、Test データを使い実験を行った。データセットの詳細を表1に示す。

表1 実験用データセット

	Positive	Neutral	Negative	Total
Train	2622	1729	719	5070
Test	7059	10342	3231	20632

単語のポジティブ性・ネガティブ性の算出には表1とは独立の表2に示す SemEval の過去データ(2013~2015年)のデータセットを利用した。単語の抽出には、NLTK³を用いた。表2のデータに関しては、NLTKの word_tokenize 関数で43253単語抽出後、出現ツイート数が3以上のものを残した9035単語を実験で用いた。

表2 ポジティブ性・ネガティブ性用データセット

	Positive	Neutral	Negative	Total
過去データ	8312	8114	2931	19357

4.2 評価指標

分類精度を評価する指標として $F_1\text{Score}(\text{pol})$ を用いる。 $F_1\text{Score}(\text{pol})$ は以下の式によって定義される。ここでは $F_1\text{Score}$ をそれぞれの極性について求め、精度を比較する。なお、極性(Polarity)を pol と表現している。

$$F_1\text{Score}(\text{pol}) = \frac{2 \cdot \text{Precision}(\text{pol}) \cdot \text{Recall}(\text{pol})}{\text{Precision}(\text{pol}) + \text{Recall}(\text{pol})}$$

4.3 実験結果

4.3.1 ポジティブ性・ネガティブ性ランキング

ポジティブ性・ネガティブ性について表2のデータを使いランキングした単語、Bi-Gram、ならびに単語ペアの上位10件をそれぞれ表3に示す。

表3 ポジティブ・ネガティブ両極性のランキング上位の単語

単語のポジティブ性の上位10件				
Fun	Awesome	Amazing	Thank	Stage
Hopefully	Congrats	Annual	Happy	Dance
単語のネガティブ性の上位10件				
Erdogan	Failed	Fucked	Syria	Isis
Worse	Clicking	Dumb	Vulnerable	Wiped
Bi-Gram のポジティブ性の上位10件				
! we	Thank you	-)	;)	Good luck
So excited	Excited for	Happy birthday	Excited to	This Saturday
Bi-Gram のネガティブ性の上位10件				
I hate	n't like	Give a	Jeb bush	, other
The gop	Of planned	Wiped by	Why does	S ii
単語ペアのポジティブ性の上位10件				
! happy	! love	, happy	Happy you	! best
And happy	Birthday happy	Excited tomorrow	Thank you	Excited so
単語ペアのネガティブ性の上位10件				
Erdogan the	Bush the	Jeb the	A palin	Erdogan of
Erdogan is	Erdogan to	@ bush	Palin she	Erdogan turkey

実験より、多くの単語がそれぞれの極性意味を持った単語となっている。ネガティブな単語ペアについては政治に関する単語が多く出現していることが確認できる。

4.3.2 重み付き平均の検討

提案システムでは提案素性と単語分散表現を用いた素性による分類結果を重み付け平均するアンサンブル手法を用いている。重み付け平均の重み w を変更し、実験した。実験では $N=600$ とした。その結果を図5に示す。

図5より提案素性($w=1$)は $F_1\text{Score}(\text{Neutral})$ が高く、GloVeによる素性($w=0$)は $F_1\text{Score}(\text{Negative})$ が高いことが分かる。また、重み付け平均によりそれぞれの手法のバランス化が図れていることが分かる。特に、 $w=0.33$ 付近で3極性とも値が高くなることが分かる。

² <https://www.csie.ntu.edu.tw/~cjlin/liblinear/>

³ <http://www.nltk.org/>

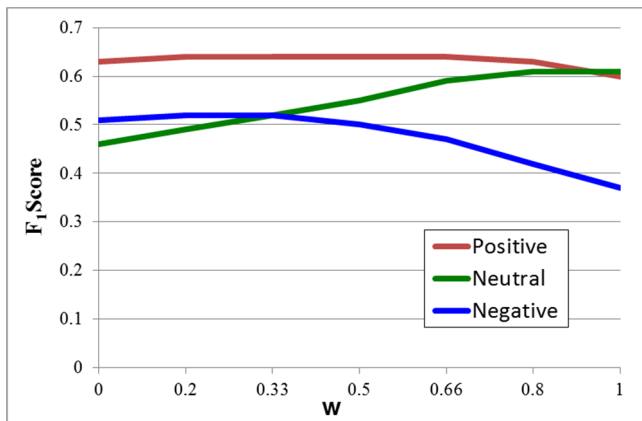


図5 W-F₁Score グラフ

4.3.3 比較実験

比較手法は SemEval2016 にて 5 位の Giorgis らの手法[2], SemEval2015 にて 1 位の Hagen らの手法[11], ならびに GloVe 単体とした. 各手法の次元数を表 4 に示す. 表 4 において, Giorgis らは, アンサンブル学習で文法的素性(1316 次元)と GloVe(200 次元)を用いている. Hagen らは, 過去参加した 4 チームの素性をアンサンブルしたものである. 本実験では N=600 とし, 重み w は 0.33 とした. 実験結果を表 5 に示す.

表 4 提案手法・比較手法の次元数

Method	次元(括弧はアンサンブルの意)
Giorgis ら	(1316, 200)
Hagen ら	(416687, 26226, 75016, 416643)
GloVe 単体	200
提案手法	(1202, 200)

表 5 比較実験結果

Method	F ₁ Score (Positive)	F ₁ Score (Neutral)	F ₁ Score (Negative)
Giorgis ら	0.629	0.500	0.469
Hagen ら	0.634	0.502	0.474
GloVe 単体	0.630	0.462	0.510
提案手法	0.641	0.518	0.522

表 4, 5 より, 従来手法に比べて, 小さい次元数で 3 極性いずれの評価指標でも上回っており, 分類精度が向上していることが分かる.

5. おわりに

本研究では, 単語のポジティブツイートとネガティブツイートでの出現比率に着目し, ポジティブ性・ネガティブ性を定義し, ツイートの 3 極性分類を行う新しい手法を提案した. 単語のポジティブ性・ネガテ

ィブ性の上位の単語を用いた素性を提案し, それを用いた分類器の訓練システム・テストシステムを構築した. その結果, 従来手法との比較実験を行い, 分類精度が向上することを確認した.

今後の課題としては, ランキングした単語の順位まで考慮したスコアリング方法の検討が挙げられる. また, SemEval2016 にて, 1-4 位を深層学習による手法が獲得したことから, 深層学習の導入による全体的な分類精度の向上も今後の課題として挙げられる.

謝辞

本研究の一部は, 科研費基盤 (B) (課題番号 26280038) の支援を受けて遂行した.

参考文献

- [1] A. Go, R. Bhayani, L. Huang, **Twitter Sentiment Classification using Distant Supervision**, *CS224N Project Report*, vol.1 12 pp., January 2009.
- [2] S. Giorgis, A. Rousas, et al., **auhb.twitter.sentiment at SemEval-2016 Task 4: A Weighted Ensemble of SVMs for Twitter Sentiment Analysis**, *SemEval 2016*, pp. 96-99, June 2016.
- [3] M. Rouvier, et al, **SENSEI-LIF at SemEval-2016 Task 4: Polarity embedding fusion for robust sentiment analysis**, *SemEval 2016*, pp. 202-208, June 2016.
- [4] J. Deriu, et al., **SwissCheese at SemEval-2016 Task 4: Sentiment Classification Using an Ensemble of Convolutional Neural Networks with Distant Supervision**, *SemEval 2016*, pp. 1124-1128, June 2016.
- [5] B. O'Connor, R. Balasubramanian, et al., **From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series**, *ICWSM vol.11*, pp. 122-129, May 2010.
- [6] G. Paltoglou, M. Thelwall, **Twitter MySpace Digg: Unsupervised Sentiment Analysis in Social Media**, *J. of ACM TIST*, 3(4), 19 pp., September 2012.
- [7] O. Almatrafi, S. Parack, B. Chavan, **Application of Location-Based Sentiment Analysis Using Twitter for Identifying Trends Towards Indian General Elections 2014**, *IMCOM 2015*, 5 pp., January 2015.
- [8] D. Buscaldi, I. Hernández-Farías, **Sentiment Analysis on Microblogs for Natural Disasters Management : A Study on the 2014 Genoa Floodings**, *WWW 2015*, pp. 1185-1188, May 2015.
- [9] J. Bollen, H. Mao, X. Zeng, **Twitter mood predicts the stock market**, *J. of Com. Sci.* 2(1), pp. 1-8, April 2011.
- [10] P. Nakov, et al., **SemEval-2016 Task 4: Sentiment Analysis in Twitter**, *SemEval 2016*, 18 pp., June 2016.
- [11] M. Hagen, M. Potthast, M. Büchner B. Stein, **Webis: An Ensemble for Twitter Sentiment Detection**, *SemEval 2015*, 8 pp., June 2015.
- [12] A. Giachanou, F. Crestani, **Like It or Not: A Survey of Twitter Sentiment Analysis Methods**, *ACM Comp. Surveys* 49(2), 42 pp., November 2016.
- [13] N. Felix et al., **A Survey and Comparative Study of Tweet Sentiment Analysis via Semi-Supervised Learning**, *ACM Comp. Surveys* 49(1), 40 pp, July 2016.