E-011

単語位置と強弱表現に着目したツイートの感情分析

Sentiment analysis of tweets focusing on the position of polarity words and on the emphasized and de-emphasized expressions

三和 未佐希† 立間 淳司‡ 青野 雅樹‡
Misaki Miwa Atsushi Tatsuma Masaki Aono

1. はじめに

Twitter では多くの人が様々なことについてツイートをしている．ある単語で検索すれば，その単語を含むツイートが表示され，その単語に向けられた意見などを知ることができ，感情の傾向を把握できる．

しかし，大量のツイートがある場合，すべてのツイートを読み，意見や感情を把握することは困難である．この問題に対して，本研究では，ツイートの感情を自動的に判定することを考えた．

2. 目的

本研究では，与えられたツイートに正負の極性値を自動的に与えることを目的とする．ツイートには様々な感情があるが，極性値が正であれば「ポジティブ」，負であれば「ネガティブ」，0 に近ければ「どちらでもない」と判断できるようにする．また，どの程度ポジティブ，ネガティブなのかを，極性値の大きさから判断できるようにする．

本研究で想定する，ツイートに対する極性値の付与の例を表１に示す．このように，ツイートの極性値を５段階に分けることを目標とする．

表１　ツイート極性値付与の例

極性値	ツイートの例
1.0	おなかいっぱい幸せ！
0.5	まあまあいい経験ができたと思う
0	今日の３時から講演会があるらしい
-0.5	楽しかったけど足がだるい気がする
-1.0	今日は疲れたもう帰りたい

3. 関連研究

ツイートの感情分析に関する研究には様々なものがあるが，例えば山内らの研究[1]では，番組関連ツイートの感情を，８つの感情に基づいて分析している．感情語に対して８つの感情極性値を与えた感情極性辞書と，係り受けルールを利用し，感情語と強調表現などを考慮して，ツイートの感情を導き出している．

また，高村らの研究[2]では，複数語で表される評価表現をポジティブ，ニュートラル，ネガティブという極性に分類している．

4. 提案手法

ツイート極性値を計算するにあたっては，まず，与えられたツイートを形態素解析した後，一部の品詞を省く．そして，二つの自作辞書を用いてツイート極性値を計算する．以降，計算に利用する形態素と辞書，および，極性値の計算手法について述べる．

†豊橋技術科学大学　大学院工学研究科　情報・知能工学専攻
‡豊橋技術科学大学　大学院工学研究科　情報・知能工学系

4.1 形態素解析

ツイートの形態素解析は，形態素解析器 lucene-gosen[3]を使用して行った．今回は IPA 辞書を内包したバージョンを利用し，辞書の拡張は行わなかった．

収集したツイートを形態素解析したあと，品詞の情報に「名詞」「動詞」「形容詞」「副詞」「接続詞」「助動詞」「記号」「フィラー」が含まれていれば，その形態素の原形を保存した．この場合に省かれる品詞は「接頭詞」「助詞」「その他」「未知語」となる．

4.2 単語極性値・単語影響値の辞書作成

2013 年 10 月 8 日に 2714 件のツイートを Streaming API を使用して収集した．本研究では，どのようなツイートが与えられても対応できるシステムを目標とするため，扱うツイートは特定の話題に関するものではなく，ランダムなものとした．

これらのツイートを「ポジティブ」「少しポジティブ」「どちらでもない」「少しネガティブ」「ネガティブ」の５段階で評価しながら，評価の際に決め手となった単語を自作辞書に追加していく作業を人手で行った．

自作辞書は「極性値のある単語の辞書」と「強弱表現などの辞書」の二つであり，前者の辞書は単語極性値，後者の辞書は単語影響値が単語に付与される．

4.2.1 極性値のある単語の辞書（極性値辞書）

極性値のある単語の辞書には「ポジティブ」「少しポジティブ」「少しネガティブ」「ネガティブ」の４つのカテゴリを設定した．

「少しポジティブ」「少しネガティブ」は，ものの状態を表すような単語や，その単語自体にはポジティブやネガティブといった意味があるとしても，一つ文中に現れただけではツイート全体の感情を大きく変えないと考えられる単語を中心に構成した．（例：お願い，違和感）

「ポジティブ」「ネガティブ」は，人の考え方，思ったこと，意見などを表すような単語を中心に構成した．（例：ありがとう，つまらない）

4.2.2 強弱表現などの辞書（影響値辞書）

強弱表現などの辞書には「強調」「補足」「弱化」「打消」の４つのカテゴリを設定した．

「強調」は，その単語が入ることにより，文に含まれる感情が強まる単語を中心に構成した．（例：とても）

「補足」は，強調ほどではないが，感情が少し強まる単語を中心に構成した．（例：（し）たい）

「弱化」は，その単語が入ることにより曖昧さが増したり，控えめな感情になったりするような単語を中心に構成した．（例：だいたい）

「打消」は，その単語が入ることにより，極性値のある単語の感情を反転させるような単語を中心に構成した．（例：ない）

二つの辞書のカテゴリに対して付与した単語極性値と単語影響値を表２，表３に示す．

表２　単語極性値b

カテゴリ	極性値
ポジティブ	1.0
少しポジティブ	0.4
少しネガティブ	-0.4
ネガティブ	-1.0

表３　単語影響値e

カテゴリ	影響値
強調	1.5
補足	1.1
弱化	0.2
否定	-0.8

4.3 ツイート極性値の計算

図１に，提案するツイート極性値の計算手法を示す．

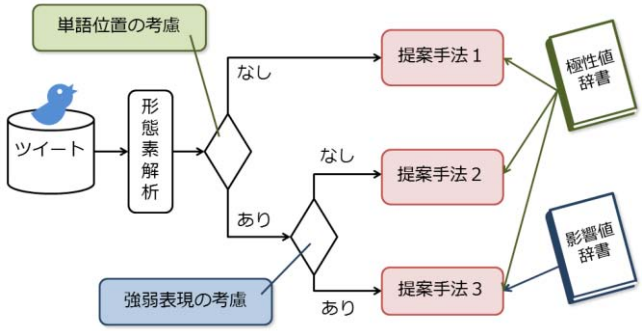

図１　ツイート極性値の計算手法

全単語数$length$，ポジティブ系単語の数c_p，ネガティブ系単語の数c_n，単語番号i，位置による単語極性値b_i，単語影響値e_iとして，ツイート極性値aを，以降の三種類の提案手法を使用して求める．

ここで，単語極性値bは，極性値辞書に含まれていれば対応した極性値，なければ 0 である．また，単語極性値eは，影響値辞書に含まれていれば対応した数値，なければ 1 である．さらに，式中のパラメータについては，$\alpha = 1$，$\beta = 0.1$，$\mu = 1$とした．

4.3.1　提案手法１（単語位置考慮なし）

ツイート中に現れた単語極性値を合計し，極性値のある単語の数を使って調整することで，ツイート極性値aを求める．

$$a = \frac{\sum_i^{length} b_i}{\alpha c + \beta\left(length - (c_p + c_n)\right)} \qquad (1)$$

$$c = \begin{cases} 1 & (c_p = c_n \text{のとき}) \\ |c_p - c_n| & (\text{それ以外}) \end{cases}$$

4.3.2　提案手法２（単語位置考慮あり）

単語位置iによってbに対する係数を変え，式(1)のb_iに代入することで，ツイート極性値aを求める．式(2)は，単語が文末に近いほど，単語極性値に対して重みを加える．

$$b_i = b \cdot \exp\left(-\left(\frac{i}{length} - \mu\right)^2\right) \qquad (2)$$

4.3.3　提案手法３（単語位置・強弱表現考慮あり）

提案手法２で求めたaを利用し，以下の式に代入することで，ツイート極性値a_eを求める．影響値をかけあわせることで，強調・補足表現が含まれるツイートの極性値の絶対値は大きくなり，弱化表現が含まれるツイートの絶対値は小さくなる．否定表現の場合は，負の数をかけあわせることで，元の値の符号を反転させる．

$$a_e = a \prod_i^{length} e_i \qquad (3)$$

5.　実験

2013 年 10 月 18 日に 310 件のツイートを Streaming API を使用して収集した．これらのツイートを人手で「ポジティブ」「少しポジティブ」「どちらでもない」「少しネガティブ」「ネガティブ」の５段階で評価し，正解データセットを作成した．

提案手法で計算したツイート極性値aまたはa_eが，-0.6 未満は「ネガティブ」，-0.6～-0.2 は「少しネガティブ」，-0.2～0.2 は「どちらでもない」，0.2～0.6 は「少しポジティブ」，0.6 以上は「ポジティブ」としてクラス分けを行い，正解データと同じクラスに入れば正解とした．結果を表４に示す．

表４　各手法の正解率

提案手法１	提案手法２	提案手法３
0.616	0.642	0.652

単語の位置を考慮した提案手法２は提案手法１より正解率が高く，強弱表現を考慮した提案手法３では，更に正解率が高くなっている．ツイートの感情分析では，単語の位置と強弱表現が重要であるといえる．

6.　まとめ

提案手法により計算したツイート極性値を用いて感情を判定すると，61%～65%の正解率となった．単語位置を考慮した方が，正解率が上がっており，さらに強弱表現を考慮することで，正解率を上げることができた．しかし，部分を否定する表現がツイート全体の否定となってしまい，誤った極性値となってしまう場合があった．また，顔文字が極性を決定しているツイートには，対応できていない．

形態素解析の際に顔文字や未知語に対応すること，構文解析を取り入れて強弱表現を考慮すること，辞書やパラメータを改良することが課題となる．

参考文献

[1] 山内 崇資, 林 佑樹, 中野 有紀子：日本語解析による Twitter の感情分析とシーンインデキシングへの応用, 情報処理学会第 75 回全国大会(2013)

[2] 高村 大也，　乾 孝司，　奥村 学：隠れ変数モデルによる複数語表現の感情極性分類(自然言語),情報処理学会論文誌 47(11), 3021-3031 (2006)

[3] lucene-gosen：https://code.google.com/p/lucene-gosen/