2D2-1

テキスト情報から生成された極性辞書を用いた市場動向分析

Market Analysis Using PolarityDictionary made from Textual Data

伊藤 友貴 *1 Tomoki Ito　坪内 孝太 *2 Tatsuo Yamashita　山下 達雄 *2 Tatsuo Yamashita　和泉 潔 *1 Kiyoshi Izumi

*1東京大学大学院工学系研究科 School of Engineering, The University of Tokyo　*2ヤフー株式会社 Yahoo! JAPAN Research

There are little research on the construction of a polarity dictionary in the form that words of similar meaning became one unit, polarity concept dictionary. In this study, we made polarity concept dictionary useful for the market trend analysis. First, we analyzed the method of constructing polarity concept dictionary, which was proposed by our previous research, from the theoretical point of view. Next, we applied the analysis to the construct of the polarity concept dictionary and improved the method. Then, using real textual datasets, we made polarity concept dictionaries and estimated the market trend. As a result of comparison with other traditional methods, the proposal method could forecast the market trend in higher F-score.

1. はじめに

情報通信技術の発達に伴い，金融テキストマイニングの技術に対する個人投資家及び機関投資家からの関心が高まってきている [1]．金融テキストマイニングとは，投資に有用な情報を SNS や記事のような大規模のテキストデータから抽出する技術である．この分野においては，極性辞書を使うことが有用である [2]．極性辞書とは単語の極性情報に関する辞書である．特定の文脈においてポジティブな意味を持つ単語には正の極性値を，ネガティブな意味を持つ単語には負の極性値を与える形で各単語にその極性値を割り振っている辞書である．

通常，極性辞書は人手によって作成されるが，経済用語・ネットスラングについては十分な極性についての情報を極性辞書からは手に入れられない．経済用語のポジネガ極性付与についての研究もいくつかされているが [3, 4]，これらの方法により作られるポジネガ辞書や現在世間に出回っている極性辞書からは，単語単位でのポジネガ情報は抽出できるものの，単語間の類似度は取ることができない．よって，異なる語群からなるテキストの情報を同じ指標による分析が難しい．このような分析には，似た意味をもつ単語がひとまとまりの概念になっている形での極性辞書，即ち極性概念辞書が必要である．著者らによって提案された II algorithm [5] は極性概念辞書構築手法の一つである．しかし，II algorithm が適切に極性を付与するかどうかに関する理論的な側面からの検証はされていない．

本研究の目的は II algorithm が適切に極性を付与する条件について理論的に解析し，II algorithm による単語への極性付与手法の妥当性を示すこと，及び 理論解析に基づいた II algorithm による極性概念辞書構築手法の開発である．

まず，II algorithm が適切に極性を付与する条件について理論解析をもとにに求めた．その後，word2vec [6] を用いてテキストに出てくる単語にベクトルを与え，それをもとに各テキストに特徴量ベクトルを与えた．これらの単語のうちの一部には経済専門家の手によってつけられた極性スコアが付与されている．極性が不明な単語の極性を求めるために，各テキストの特徴量とその極性タグの対応の関係をニューラルネットワークモデルを用いて分析した．ニューラルネットワークの設計には理論解析から得られた知見が活かされている．機械学習の学習の過程で極性辞書に含まれる単語の極性スコアが極性辞書外の単語にも伝播することが期待できる．本手法によって得られる特徴量を用いた極性タグの予測結果と既存手法から得られる特徴量を用いた予測結果を比較し，II algorithm が与える極性値の妥当性を検証した．最後に，ロイターニュースとヤフーファイナンス掲示板から極性概念辞書を作成した．

連絡先: 伊藤友貴，M2015titoh@socsim.org

2. 極性概念辞書構築手法

本節では単語への極性値付与手法 II (importance infiltration) algorithm [5] を紹介する．

2.1 Word classification and document representation 法 (CDR 法)

まず，word2vec によって単語に与えた分散表現を利用した文書の特徴量の生成手法，CDR 法 [7] の紹介をする．CDR 法は提案手法 II algorithm のベースとなった手法である．CDR 法では，意味の近い単語が同じクラスタになるようにクラスタリングし，文書内に出現する各クラスの出現回数によって文書の特徴量を生成する．まず，word2vec [6] を使い，ニュース記事に出現する各単語にベクトル表現を与える．その後，クラスタ数 K を決めた後，K-means 法 [8] により単語のクラスタリングを行い，各文書の特徴量 $\boldsymbol{V}_{\text{document}}(\in \mathbb{R}^K)$ を文書中に出現する各クラスタの回数により算出する手法である．

2.2 Importance Infiltration propagation algorithm

word2vec で単語をクラスタリングすると，対義語同士が同じクラスに入ってしまうことがある．これは word2vec では各単語がどの言葉と組み合わせて使われるかによって単語の分散表現を獲得するためである．しかし，これは市場動向の分析をする上では望ましくない．そこで，本研究ではこれから紹介する II (importance infiltration) algorithm) を用いて極性辞書外単語に極性を与え，文書のベクトルを生成した．文書のベクトルを生成するにあたり，金融機関に所属する機関投資家が人手で作成した経済用語極性辞書の情報を用いた．II algorithm では 図 1 のように表現されるニューラルネットワークモデルを用いた．入力層の次元は単語の数で、各単語に対応するノード (1 層目の各ノード) は、それぞれ属するクラスタを表すノード (2 層目のノード) のみに結合するという構造のニューラル

ネットワークモデルである．ここで，$\boldsymbol{W}_{\text{polarity}} \in \mathbb{R}^{K\times m}$ (式 (4) にて定義, $\boldsymbol{W}_3 \in \mathbb{R}^{2\times K}$ は重み行列，$b_0 \in \mathbb{R}^2$ はバイアスベクトル , $y_{cls} \in \{0\ (\text{ネガティブ}), 1\ (\text{ポジティブ})\}$ は出力層の値である．y_{cls} は文書につけられるタグに該当する値である．文書番号 j の文書内に出現する単語の頻度からなるベクトル $\boldsymbol{V}_{\text{BOW}j}$ を式 (1) のように定義する．

$$\boldsymbol{V}_{\text{BOW}j} := \boldsymbol{Z}_j^{(1)} = [\boldsymbol{Z}_{j,1}^{(1)^T}, \boldsymbol{Z}_{j,2}^{(1)^T}, \cdots, \boldsymbol{Z}_{j,K}^{(1)^T}]^T \quad (1)$$

$\boldsymbol{Z}^{(l)}$ は l 層目の出力を意味する．$\boldsymbol{Z}_{j,k}^{(1)} \in \mathbb{R}^{n(k)}$, $\boldsymbol{Z}_{j,k}^{(1)}[i](k = 1, 2, \cdots, K)$ は 単語 $w_{k,i}$ ($w_{k,i}$ は単語のクラスが k でクラスタ内単語 ID が i の単語とする．) の文書番号 j の文書における頻度， $n(k)$ はクラス k である単語の数である．このとき，II algorithm で用いるニューラルネットワークモデル (図 1) は式 (2), 式 (3) のように y_{cls} を表すことで表現できる．

$$\boldsymbol{y_j} = f_3(\boldsymbol{W}_3(\tanh(\boldsymbol{W}_{\text{polarity}}\boldsymbol{V}_{\text{BOW}j}) + \boldsymbol{b_0}) \quad (2)$$

$$y_{cls_j} = \text{argmax}\ \boldsymbol{y_j} \quad (3)$$

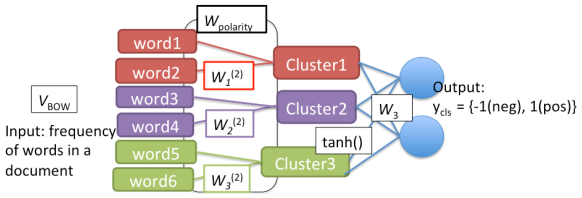

図 1: 本研究で用いたニューラルネットワークモデル (3 層)

活性化関数 f_3 には Softmax 関数を用いた．また，学習時の損失関数には Softmax cross entropy 関数を用いた．過学習を防ぐために学習時には, Dropout 法 [9] を用いた．$\boldsymbol{W}_{\text{polarity}}(\in \mathbb{R}^{K\times m})$ を以下のように定義する．

$$\boldsymbol{W}_{\text{polarity}} := \text{diag}(\boldsymbol{W}_1^{(2)^T}, \boldsymbol{W}_2^{(2)^T}, \cdots, \boldsymbol{W}_K^{(2)^T}) \quad (4)$$

$\boldsymbol{W}_k^{(2)} \in \mathbb{R}^{n(k)}$ であり，$\boldsymbol{W}_k^{(2)}[i]$ は $w_{k,i}$ の極性値に対応する．

次に $W^{(2)}$ の初期値の与え方について説明する．単語 $w_{k,i}$ が既存の極性辞書内の単語であり，極性値が事前に専門家の手によって与えられている場合, 単語 $w_{k,i}$ の極性辞書値 $PS(w_{k,i})$ を用いて $\boldsymbol{W}_k^{(2)}[i]$ の初期値を以下のように与える．

$$\boldsymbol{W}_k^{(2)}[i] = \begin{cases} PS(w_{k,i}) & (\text{単語 } w_{k,i}\text{が極性辞書単語のとき}) \\ 0 & \text{それ以外} \end{cases}$$

とする．教師あり学習の過程で, $\boldsymbol{W}_{\text{polarity}}$ の値が更新されていく． 学習後に $\boldsymbol{W}_{\text{polarity}}$ の値を取り出すことで既存極性辞書外単語の極性値を得ることができる．$\boldsymbol{W}_k^{(2)}[i]$ の値が伝搬後の単語 $w_{k,i}$ の極性値に対応する．

次に理論解析から判明した II algorithm の性質を述べる．

2.3 定義

学習時のミニバッチサイズ N を用いて入力値 $\boldsymbol{X}(\in \mathbb{R}^{m\times N})$, j^+, j^-, p^+ , p^-, $\boldsymbol{U}$, $\boldsymbol{Z}$ を以下に定義する．

$\boldsymbol{X} := [\boldsymbol{V}_{\text{BOW}1}, \boldsymbol{V}_{\text{BOW}2}, \cdots, \boldsymbol{V}_{\text{BOW}N}]$

$p^-(w_{k,i}) := p(w_{k,i}\text{出現文書がネガティブ}\ |w_{k,i}\text{が文書に出現})$

$p^+(w_{k,i}) := p(w_{k,i}\text{出現文書がポジティブ}\ |w_{k,i}\text{が文書に出現})$

$$\boldsymbol{U}^{(2)} := \boldsymbol{W}_{\text{polarity}}\boldsymbol{X}, \boldsymbol{Z}^{(2)} := \tanh(\boldsymbol{U}^{(2)})$$

$$\boldsymbol{U}^{(3)} := \boldsymbol{W}_3(\boldsymbol{Z}^{(2)} + \boldsymbol{b_0})$$

便宜的に j を文書番号 j の文書がポジティブである場合に j^+, 文書番号 j の文書がネガティブである場合に j^- と表記する．また，w_{k,i^+}, w_{k,i^-}， $w_{k,i^{neu}}$ を

$$w_{k,i} = \begin{cases} w_{k,i^+} & p^+_{w_{k,i}} > p^{t+} \\ w_{k,i^-} & p^-_{w_{k,i}} > p^{t-} \\ w_{k,i^{neu}} & \text{otherwise} \end{cases}$$

と定め,

$n^+(k)$: クラス k 内の単語で $p^+_{w_{k,i}} > p^{t+}$ である単語の数,

$n^-(k)$: クラス k 内の単語で $p^-_{w_{k,i}} > p^{t-}$ である単語の数,

$n^{neu}(k)$: 上記二つの条件を満たさない単語の数

と定める．

2.4 極性伝搬条件

II algorithm について以下に述べる単語への極性付与に関する性質が成り立つ．

(i) ポジティブ単語 w_{k,i^+} に十分大きい正の極性値,
ネガティブ単語 w_{k,i^-} に十分小さい負の極性値が付与

(ii) $p^-(w_{k,i^-})$, $p^+(w_{k,i^+})$, $n^-(k)$, $n^+(k)$ が十分に大きい

(iii) $\boldsymbol{W}_3[0][k] < 0$ 十分に小さく, $\boldsymbol{W}_3[1][k]$ 十分に大きい

(iv) ミニバッチサイズ N が十分に大きい

このとき, 任意の i^-, i^{neu}, i^+ について条件 (i)–(iv) を満たすように初期値が与えられる場合には

$$E[\partial \boldsymbol{W}_k^{(2)}[i^+]] < 0, E[\partial \boldsymbol{W}_k^{(2)}[i^-]] > 0$$

となり，ポジティブ単語 w_{k,i^+} には正の極性値が与えられ，ネガティブ単語 w_{k,i^-} には負の極性値が与えられる．これより II algorithm による単語への極性付与が妥当だと言える．

本主張は条件 (i)–(iv) が成り立つという条件下での誤差逆伝搬の様子を計算して求めることで示すことができるが，具体的な説明については省略する．本主張より，条件 (iii) を満たすように $\boldsymbol{W}_3$ の初期値を与えた上で II algorithm により各単語へ極性値を付与することで $\boldsymbol{W}_3$ の値を乱数によって決める場合に比べ妥当に極性値を付与できると考えられる．

3. 実データを用いた極性伝搬の検証

実データを用いた検証により,「II algorithm が付与する単語の極性値の有用性」を検証した．有用性を調査するにあたり以下の実験をした．

3.1 株価動向分析

本実験においては 2013 年 1 月から 2015 年 12 月までの間に配信されたトムソンロイターの経済ニュース記事のうち．銘柄コードの入りの記事を用いた．これらの記事が個別銘柄の株価動向に与える影響を予測した．まず，ニュース記事の配信日を d, 記事に最初に銘柄コードが出てくる個別銘柄の日付 $d-1$ の終値 p_{t_p}, 及び日付 $d+1$ の始値 p_{t_a} から算出される株価リターンを用いて以下の y_{cls} で定めるタグを各記事につけた．

$$\Delta p(d) = \frac{p_{t_\text{a}} - p_{t_\text{p}}}{p_{t_\text{p}}}, y_{\text{cls}} = \begin{cases} 1 & (\Delta p(d) > 0.01) \\ -1 & (\Delta p(d) < -0.01) \end{cases}$$

その後，記事のタグの予測を行い，F 値の 10 交差検定平均スコアを算出し評価した．

3.2 ヤフーファイナンス掲示板ポジネガタグ予測

Yahoo!Finance 掲示板 [*1] の投稿に付与されるタグの予測をおこなった．タグの種類は $\{1,2,3,4,5\}$ である．1 は「強く買いたい」, 5 は「強く売りたい」を意味する．以下の二つのケースについて実験を行った．

3.2.1 短期間における多種多様な銘柄の投稿に関する予測

2014/9/1 ～ 2014/9/30 までの全銘柄の投稿について投稿日の古いタグ 1 とタグ 5 の投稿 10000 件ずつを抽出し，5 交差検定ルールに基づいて訓練用データと検証用データに分け予測し，F 値の 5 交差検定平均スコアをもとに評価した．

3.2.2 長期間における少数銘柄の投稿に関する予測

まず，2014/11/18 ～ 2016/6/15 まで銘柄番号 998407, 9501, 4777, 8462, 4564, 6871 のタグが 1 または 5 の投稿を抽出した．2015/6/1 ～ 2016/5/31 について各月の投稿を検証データ，各月の前月以前のタグが 1 と 5 の投稿各 20000 件ずつを訓練用データとして予測精度を検証した．月毎に F 値を算出し，12 回の平均スコアをもとに評価した．

3.3 前処理

本実験において，形態素解析には MeCab [11] を用いた．このとき，形態素の並びが「動詞 + 助動詞 (未然形)」,「動詞 + 動詞」,「動詞 + 助動詞 + 助動詞 (未然形)」の場合は一語にまとめ動詞とした．また，3 回以上繰り返す文字列は 2 回に圧縮するなどして，各単語の正規化をおこなった．実験 3.1 節では動詞，形容詞，名詞を抽出 (全部で 21286 語) し，また実験 3.2.1 項では動詞，形容詞，名詞・未知語を抽出 (全部で 25218 語) し，各単語の頻度から文書の特徴量を生成した．実験 3.2.2 項では形態素を抽出した上で，文書全体にて 100 回以上出現する極大部分形態素列 [12] 16627 表現を抽出し，それらの頻度をもとに文書の特徴量を生成した．

3.4 ベースライン

Label [13], SSL [4] ，Ridge Regression [3] の各既存手法及び, II algorithm (3 層) (II algorithm (layer3)) によって得られた各単語の極性値 ($PS^*(word_l)$ を単語 $word_l$ とする) と CDR 法 (2.1 節) を用いて以下のように $\boldsymbol{W}_{\text{polaritybase}}$ を生成し，$\boldsymbol{V}_{\text{BOW}}\boldsymbol{W}_{\text{polaritybase}}$ の値を文書の特徴量とした．

$$\boldsymbol{W}_{\text{polaritybase}} = (\delta^*_{lj})^T$$

$$\delta^*_{lj} = \begin{cases} PS^*(word_l) & (word_l \in class_j) \\ 0 & \text{else} \end{cases}$$

これらによる予測結果，及び CDR 法 (任意の l について $PS^*(word_l) = 1$ である場合に相当)，AntSyn [14] を用いた CDR 法 (AntSyn)，$\boldsymbol{V}_{\text{BOW}}$ (BOW) により生成された特徴量による予測結果の間で比較した．予測モデルには線形 SVM を用いた．II algorithm に利用した経済用語極性辞書の単語数は約 200 であった．II algorithm の精度は, 5 回の試行平均値をもとに算出した．$K = 500$ とした．また，2.4 節の結果を踏まえ，$\boldsymbol{W}_3$ の初期値は $\boldsymbol{W}_3[0][k] \sim U(-0.01, 0), \boldsymbol{W}_3[1][k] \sim U(0, 0.01)$ に従って与えた．3.2 節における epoch 数 (学習回数) は訓練データ内におけるチューニングにより決定し，3.1 節では epoch 数 (学習回数) を 50 とした．

3.5 実験結果

表 1 が株価動向分析 (3.1) の結果である．表 2 がヤフーファイナンス掲示板ポジネガタグ予測 (3.2 節) の結果である．この結果より，II algorithm が与える単語の極性値から生成される特徴量から作られる予測モデルの方が他の手法から生成される特徴量により作られる予測モデルに比べて F 値を指標とした場合に 高い予測力を持つことがわかった．

表 1: 株価動向分析の結果

Methods	F 値 ($\pm\sigma$)
BOW	0.586
CDR	0.538
AntSyn	0.542
Ridge Regression	0.578
SSL	0.545
Label	0.525
II algorithm (layer3)	0.608 ($\pm$ 0.005)

表 2: ヤフーデータ掲示板ポジネガタグ予測

Methods	F 値 ($\pm\sigma$) (短期)	F 値 ($\pm\sigma$) (長期)
BOW	0.735	0.790
CDR	0.684	0.724
Ridge Regression	0.754	0.786
AntSyn	0.684	0.700
SSL	0.570	0.660
Label	0.633	0.651
II algorithm (layer3)	0.763 ($\pm$ 0.001)	0.795 ($\pm$ 0.001

3.6 極性概念辞書出力結果

最後に，本実験によって作成された極性概念辞書の一部を紹介する．表 3–4 はロイターニュースをもとに II algorithm によって作成した極性概念辞書，及びヤフーファイナンス掲示板から作成した極性概念辞書の結果の一部である．例えば，表 3 から「上昇」のプラスの極性値が II algorithm によって同じクラスの「急騰」や「急上昇」に伝搬し，逆に「下落」や「下降」のマイナスの極性値が II algorithm によって同じクラスの「急降下」や「暴落」に伝搬している様子がわかる．

4. まとめ

本研究において，極性概念辞書構築手法 II algorithm の性質を理論的に解析し，II algorithm が適切に各単語に極性付与する条件を求め，その結果をもとに II algorithm による極性概念辞書構築手法を構築した．また，本研究における解析によって改良された II algorithm が市場動向分析において既存の単語極性付与手法に比べ有用であることを実データによる実験によって示した．今後の課題として，極性概念辞書を用いたより踏み込んだ市場動向の分析などが考えられる．

参考文献

[1] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, D. C. L. Ngo, "Text mining for market prediction: A systematic review", Journal of Expert Systems with Applications, Vol. 41, Issue 16, pp. 7653–7670, 2014

*1 http://textream.yahoo.co.jp/category/1834773

表 3: ヤフーファイナンス掲示板から作成された極性概念辞書

単語	極性値 (伝搬後)	(極性辞書値)
上昇	0.328	(0.5)
反発	0.232	(0.5)
動意	0.297	(0.0)
動き	0.276	(0.0)
急落	-0.525	(0.0)
反転	0.498	(0.0)
下落	-0.655	(-0.333)
下降	-0.818	(-0.333)
急騰	0.184	
下げ	-0.307	
短期間	0.007	
急上昇	0.278	
反騰	-0.015	
上下	0.137	
値動き	0.095	
急降下	-0.411	
上げ	0.017	
調整	0.366	
盛り上がり	0.096	
乱高下	0.081	
上下動	0.333	
上がり	0.208	
リバウンド	-0.240	
上げ過ぎる	-0.160	
暴落	-0.791	
暴騰	0.245	
連騰	0.240	
相場	0.083	
上げ下げ	0.245	

表 4: ロイターニュースから作成した極性概念辞書

単語	極性値 (伝搬後)	(極性辞書値)
続伸	1.357	(1.118)
急伸	0.678	(0.5)
反発	0.594	(0.5)
上昇	0.443	(0.5)
急落	-0.070	(0.0)
推移	0.167	(0.0)
下落	-0.453	(-0.333)
低下	-0.384	(-0.375)
続落	-1.229	(-0.929)
反落	-1.250	(-0.938)
急騰	-0.088	
下げ	-0.120	
下押し	-0.096	
下押す	0.118	
鬼門	0.072	
下がる	-0.052	
伸び悩む	-0.034	
軟化	-0.171	
値上がり	0.025	
値下がり	0.060	
安	0.069	
買われる	0.060	
乱高下	0.261	
安い	0.155	
下げる	-0.178	
急上昇	0.089	
上げ	-0.160	
下げ止まる	0.200	
ドル安	-0.033	
伸し	0.082	

[2] W. Ye and F. Ren, "Learning sentimental influence in twitter", ICFCSA, 2011

[3] K. Tsubouchi and T. Yamashita, "Positive / Negative Detection for Finance Contents via Stock Bulletin Boards Data", JSAI 2014, 2014

[4] H. Yanagimoto, "Improvement of Sentiment Dictionary Using Neural Network Language Model", JSAI 2014, 2014

[5] T. Ito, K. Izumi, K. Tsubouchi, T. Yamashita, "Polarity propagation of financial terms for market trend analyses using news articles", CEC 2016, 2016

[6] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, "Distributed Representations of Words and Phrases and their Compositionality", NIPS 2013, pp. 3111–3119, 2013

[7] Y. Yuan, L. He, L. Peng, Z. Huang, "A New Study Based on Word2vec and Cluster for Document Categorization", Journal of Computational Information Systems, Vol. 10, Issue 21, pp. 9301–9308, 2014

[8] K. Hornik, I. Feinerer, M. Kober, C. Buchta, "Spherical k-Means Clustering", Journal of Statistical Software, Vol. 50, Issue 10, pp. 1–22, 2012

[9] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting", Journal of Machine Learning Research, Vol. 15, pp. 1929–1958, 2014

[10] D. P. Kingma, J. L. Ba, "Adam: A Method for Stochastic Optimization", arXiv:1412.6980, 2014

[11] T. Kudo, K. Yamamoto, Y. Matsumoto, "Applying Conditional Random Fields to Japanese Morphological Analysis", EMNLP 2004, pp. 230–237, 2004

[12] D. Okanohara,J. Tsujii, "The Categorization with all Substring Features", SDM 2009, pp. 838–846, 2010

[13] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, B. Scholkopf, "Learning with Local and Global Consistency", NIPS 2003, pp. 321–328, 2003

[14] K. A. Nguyen, S. S. im Walde, N. T. Vu, Integrating Distributional Lexical Contrast into Word Embeddings for Antonym-Synonym Distinction, ACL 2016, 2016