3L3-OS-16a-1in1

ニュース記事を用いた経済専門用語のクラスタリングと極性付与

Clustering financial terms and giving them their polarity scores using news articles

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　*2Yahoo! JAPAN 研究所 Yahoo! JAPAN Research

In the previous research, a new approach for giving the financial terms whose positive-negative polarity score is unknown, and making the feature vector of a document useful for predicting stock price trends was proposed. Our subject is to evaluate the usefulness of this approach in predicting. First, we assigned a numerical vector to a word appeared in financial news documents, and defined the feature vectors of documents. Then, we analyzed the stock price trend and the sentiment score which can be evaluated from the textual data of Yahoo! finance board using this feature vector. As a result of comparison with other traditional methods, the proposal method could forecast in higher accuracy about the stock price trends and the sentiment scores.

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，ポジネガ極性辞書を使うことが有用であることがわかっている [2].

特に，経済用語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という問題がある．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 [3, 4] が，今のところ確立された手法はない.

本研究の目的は先行研究 [5] にて提案されている，ポジネガスコアの不明な経済用語に対してポジネガスコアを割り当て株価動向の予測に有用なニュース記事の特徴量を生成する手法が，過去のデータをもとに未来の株価動向・及び極性スコアの予測を行う際に有用であるかを調査すること，及びハイパーパラメーター変更時における先行研究 [5] にて提案されている手法の予測精度を調査することである．まず，word2vec [6, 7] を用いてニュース記事に出てくる単語にベクトルを与え，それをもとに各記事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．その後，極性が不明な単語の極性を求めるために，各ニュース記事の特徴量と株価動向，及び Yahoo!Finance 掲示板上の投稿データから得られる極性スコアの対応の関係を分析した．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．先行研究 [5] にて提案されている手法によって得られる特徴量を用いた株価動向・極性スコアの予測結果と既存手法によって得られる特徴量を用いた予測結果を比較することで，先行研究 [5] にて提案されている手法によって得られる単語の極性スコアが妥当であるかどうかを検証した.

連絡先: 伊藤友貴，M2015titoh@socsim.org

2. 手法

2.1 word classification and document representation 法

まず，word classification and document representation 法 [8] を紹介する．基本的な考え方は，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，文書中に出現する各クラスの単語の回数によって文書の特徴量を生成するという考え方である．まず，word2vec [6, 7] 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た．その後，クラス数 K を決めた後，K-means 法によりクラスタリングを行った．ここで，各単語間の距離にはコサイン距離を用いた．K 個のクラスターを得た後, 各経済ニュース記事の特徴量 $V_{document}$ を文書中に出現する各クラスの単語の回数を用いて求めた．ここで，

$$V_{\mathrm{BOW}} = (f_1, f_2, \cdots, f_i, \cdots, f_m)^T \tag{1}$$

(f_i は $word_i$ の文書における出現頻度, m 文書内の総単語数である．)
，及び

$$W = (\delta_{ij})^T$$

$$\delta_{ij} = \begin{cases} 1 & \text{when} \quad word_i \in class_j \\ 0 & \text{else} \end{cases}$$

を用いると, 文書ベクトル $V_{\mathrm{document}} (\in \mathbb{R}^K)$ は次のように表現できる.

$$V_{\mathrm{document}} = W \cdot V_{\mathrm{BOW}} \tag{2}$$

2.2 Importance infiltration アルゴリズム

上のような手法で単語をクラスタリングすると，反対の意味の単語が同じクラスに入ってしまうことがある．これは word2vec が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．例えば，急騰と急落という単語はよく同じクラスに分類される．しかし，このようなことは株価動向の予測をする上では望ましくない．そこで，本論文ではニューラルネットワークモデルを利用したモデル (II (importance infiltration) アルゴリズム) [5] を用いて各単語に極性を与え，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，金融専門家の手によって作られた

極性辞書の情報を用いた．II アルゴリズム では 図 1，及び式 (3)，(4) 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． ここで，$W_1 \in \mathbb{R}^{K\times k}$． ($k$ は隠れ層の次元)，$W_2 \in \mathbb{R}^{k\times 3}$ は重み行列，$b_1 \in \mathbb{R}^k$，$b_2 \in \mathbb{R}^3$ はバイアスベクトル，$y_{cls} \in \{-1\ (\text{fall}), 0\ (\text{sideway}), +1\ (\text{rise})\}$ は出力層の値である．例えば，株価動向の予測をする場合，”−1” 文書に関連する銘柄に株価の下落を指す．同様に，”0” は横ばい，and ”+1” は上昇を指す．

$$V_{\text{category}} = \frac{W_{\text{polarity}} V_{\text{BOW}}}{\|W_{\text{polarity}} V_{\text{BOW}}\|} \tag{3}$$

$$y_{cls} = f_3(W_2 f_2(W_1(f_1(V_{\text{category}} + b_0) + b_1) + b_2) \tag{4}$$

ここで，活性化関数 f_1，f_2 には tanh を，活性化関数 f_3 にはソフトマックス関数を用いた．また，損失関数として Softmax cross entropy を用いた．さらに，過学習を防ぐために dropout 法 [9] を用いた．

また， $W_{\text{polarity}} (\in \mathbb{R}^{m\times c})$ の初期値は以下のようにした．

$$W_{\text{polarity}} = W_{i,j} \cdot (W_{\text{polaritydic}} + W_{\text{noize}}) \tag{5}$$

ここで，$W_{\text{polaritydic}}$ 及び W_{noize} は次のように定義される．

$$W_{\text{polaritydic}} = \text{diag}(PS_1, PS_2, \cdots, PS_i, \cdots, PS_m) \tag{6}$$

$word_i$ の極性辞書スコアが事前に機関投資家の手によって与えられている場合，

$$PS_i = \text{polarityscore}(word_i)$$

(polarityscore($word_i$) は $word_i$ の人手によって与えられる極性辞書スコアである．)，
とし，与えられていない場合は，

$$PS_i = 0$$

とした．また，

$$W_{\text{noize}} = \text{diag}(u_1, u_2, \cdots, u_i, \cdots, u_m) \tag{7}$$

とした．ここで，

$$u_i \in U(-\lambda, \lambda), \lambda \in \mathcal{R} \tag{8}$$

とした．学習の中で，W_{polarity} は変化する．学習後 W_{polarity} の値を見ることで，辞書にない単語の極性辞書を獲得することができた．その後 V_{category}，または $f_1(V_{\text{category}} + b_0)$ を計算することで文書のベクトル表現を求めた．これらは 極性辞書のスコア伝播の影響を受けた値である．

3. 実験

3.1 評価方法

以下の二つの実験において単語への極性付与，及び付与した極性の妥当性の検証を予測の精度を求めることで行った．

3.1.1 株価動向分析

本実験において 2013 年 1 月から 2014 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を用いた．そのうち．銘柄コードの入っている記事を用いた．これらの記事が個別銘柄の株価動向に与える影響の予測を行った．まず，ニュース記事の配信日を d，その日の関連する個別銘柄の株価

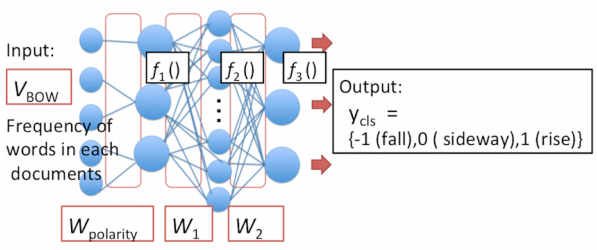

図 1: Neural NetWork Model used in this paper

を $p_{\text{code}}(d)$ とするとき，各記事に対してタグ $y_{cls} \in \{-1, 0, 1\}$ を以下のように付与した．
関連する株価のリターンが

$$\Delta p_{\text{code}}(d) = \frac{p_{\text{code}}(d+1) - p_{\text{code}}(d-1)}{p_{\text{code}}(d-1)}$$

のとき，

$$y_{\text{cls}} = \begin{cases} 1 & (\Delta p_{\text{code}}(d) > 0.01) \\ 0 & (-0.01 \le \Delta p_{\text{code}}(d) \le 0.01) \\ -1 & (\Delta p_{\text{code}}(d) < -0.01) \end{cases}$$

というルールでタグをつけた．その後，全 1801 の記事データに対して各記事を時系列順に並べ，前半の 1621 記事を訓練用データに，後半の 180 データを検証用データとして用い，訓練用データによって予測モデルを作成し，検証用データのタグを予測した．また，II アルゴリズム利用時における，クラスター数 K と隠れ層の次元数 k については，$K = k$ という条件で実験をおこなった．

3.1.2 極性予測

同じ文書を用いて，Yahoo!Finance 掲示板 [*1] の投稿から得られる個別銘柄の極性スコアの予測も行った．掲示板の極性スコアは $ysent(d, code, i) \in \{1, 2, 3, 4, 5\}$ である， ここで，d は投稿日，$code$ は投稿の銘柄コード，i は投稿の ID である．$ysent = 1$ は 「強く買いたい」，$ysent = 3$ は 「様子見」，$ysent = 5$ は 「強く売りたい」を意味する．本実験では，あるニュース記事の配信日を d とするとき，その記事のタグ $y_{cls}(d, code) \in \{-1, 0, 1\}$ を 投稿日 $d+1$ の銘柄コード $code$ に対する投稿の極性スコアの平均値を用いて以下のように付与した．

$$ysent_{\text{average}}(d, code) = \frac{1}{N} \sum_{i}^{N} ysent(d, code, i),$$

とするとき，

$$y_{cls}(d, code) = \begin{cases} 1 & (ysent_{\text{average}}(d+1, code) \ge 4) \\ 0 & (3 \le ysent_{\text{average}}(d+1, code) < 4) \\ -1 & (otherwise) \end{cases}$$

とする．ここで，N 銘柄コードが $code$ である日付 $d+1$ における投稿の総数である．全 973 の記事データに対して各記事を時系列順に並べ，前半の 876 記事を訓練用データに，後半の 97 記事を検証用データとして用い，訓練用データによって予測モデルを作成し，検証用データのタグを予測した．また，II アルゴリズム利用時における，クラスター数 K と隠れ層の次元数 k については，$K = k$ という条件で実験をおこなった．

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

表 1: Return prediction, the mean accuracy and F_1-measure on the test data set

Methods	Accuracy ($\pm\sigma$)	F_1-measure ($\pm\sigma$)
LDA (500)	0.3556	0.3509
LDA (200)	0.3222	0.3185
LDA (100)	0.3167	0.2972
LDA (50)	0.3444	0.3395
BOW	0.3444	0.3350
CDR(K = 500)	0.3556	0.3388
CDR(K = 200)	0.3722	0.3379
CDR(K = 100)	0.3056	0.2806
CDR(K = 50)	0.3667	0.3547
V_{category} (K = 500)	0.3833(±0.0243)	0.3835(±0.0255)
$f_1(V_{\text{category}} + b_0)$ (K = 500)	0.3800(±0.0255)	0.3801 (±0.0281)
V_{category} (K = 200)	**0.4011**(±0.0275)	**0.4014**(±0.0325)
$f_1(V_{\text{category}} + b_0)$ (K = 200)	**0.4022**(±0.0249)	**0.4022**(±0.0306)
V_{category} (K = 100)	0.3711(±0.0197)	0.3717(±0.0201)
$f_1(V_{\text{category}} + b_0)$(K = 100)	0.3767(±0.0419)	0.3770 (±0.0421)
V_{category}(K = 50)	0.3922(±0.0152)	0.3912(±0.0156)
$f_1(V_{\text{category}} + b_0)$ (K = 50)	0.3900(±0.0254)	0.3900 (±0.0248)

表 2: Sentiment prediction, the mean accuracy and F_1-measure on the test data set

Methods	Accuracy ($\pm\sigma$)	F_1-measure ($\pm\sigma$)
LDA (500)	0.4021	0.3998
LDA (200)	0.3814	0.3772
LDA (100)	0.3299	0.3221
LDA (50)	0.3041	0.3093
BOW	0.3711	0.3735
CDR(500)	0.3402	0.3236
CDR(200)	0.3608	0.3630
CDR(100)	0.2784	0.2827
CDR (50)	0.3299	0.3273
V_{category}(K = 500)	0.4000(±0.0335)	0.3992 (±0.0335)
$f_1(V_{\text{category}} + b_0)(K = 500)$	0.3938(±0.0476)	0.3936(±0.0472)
V_{category}(K = 200)	0.4268(±0.0654)	0.4273(±0.0652)
$f_1(V_{\text{category}} + b_0)(K = 200)$	0.3959(±0.0556)	0.3972 (±0.0550)
V_{category}(K = 100)	0.4268(±0.0082)	0.4301 (±0.0118)
$f_1(V_{\text{category}} + b_0)(K = 100)$	0.4165(±0.0202)	0.4191(±0.0246)
V_{category}(K = 50)	**0.4412**(±0.0308)	**0.4408** (±0.0296)
$f_1(V_{\text{category}} + b_0)(K = 50)$	**0.4289**(±0.0191)	**0.4280**(±0.0183)

3.2 ベースライン及び評価基準

本実験において，形態素解析は Mecab [10] を用いた．ここで，ユーザー辞書として，日経シソーラスのリスト *2 , ウィキペディアのタイトル *3, はてなのタイトルリスト *4 ニコニコ大百科のタイトルリスト *5 を用いた．また，名詞，形容詞，形容動詞，助詞を用いて各文書の特徴量を生成した．

また， II アルゴリズム によって生成される単語のポジネガ極性の妥当性を調べるために, II アルゴリズム によって生成される V_{category}, 及び $f_1(V_{\text{category}}+b_0)$ を用いて予測した結果と LDA [11], bag of words(BOW), the word classification and document representation method (CDR) によって生成される特徴量による予測結果と比較をおこなった．評価指標として，accuracy，f-1 measure を用いた．また，予測モデルには Linear SVM を用いた．なお, V_{category}, 及び $f_1(V_{\text{category}}+b_0)$ を用いた予測結果に関しては，それぞれ各 5 回ずつの数値実験を繰り返し，それにより得られた値 (accuracy, f-1 measure) の平均値を求めることによって求めた．

3.3 結果

3.3.1 実験設定

本実験において， λ (式 (8) にて定義) の値は 0.01 とした．表 1，表 2 それぞれ実験 3.1.1，実験 3.1.2 の結果である．表 1, 表 2 の結果から, クラスタリングにおけるクラスタ数を変えた場合においても，II アルゴリズムによって生成される特徴量 V_{category} ， $f1(V_{\text{category}} + b_0)$ が他の既存手法によって生成される特徴量よりも株価動向の予測，及び極性スコアの予測の上で有用であることが確認できた．

4. まとめ

本論文において，先行研究 [5] にて考案されている手法が株価動向予測・掲示板投稿における感情極性スコアの予測をする場合において他の既存手法に比べ有用であることを実験的に示すことができた．今後の課題として，より大規模なデータにおける実験，より安定的で高い予測精度を出すための次元削減の手法の考案，初期値パラメーターの割り当て方なども含めた予測モデル構築法の考案が考えられる．

*2 The name list can be downloaded from `http://t21.nikkei.co.jp/public/help/contract/price/01/help_kiji_thes_field.html`

*3 The title list can be downloaded from `http://dumps.wikimedia.org/jawiki/latest/jawiki-latest-all-titles-in-ns0.gz`

*4 The name list can be downloaded from `http://d.hatena.ne.jp/images/keyword/keywordlist_furigana.csv`

*5 The name list can be downloaded from `http://www.nii.ac.jp/cscenter/idr/nico/nicopedia-apply.html`

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", *Expert Systems with Applications, Volume 41, Issue 16, 15 November 2014*, pp. 7653-7670, 2014

[2] W. Ye and F. Ren, "Learning sentimental influence in twitter." *Future Computer Sciences and Application(ICFCSA), 2011 International Conference on. IEEE*, 2011.

[3] K. Tsubouchi and T. Yamashita, "Positive / Negative Detection for Finance Contents via Stock Bulletin Boards Data", *The 28th Annual Conference of the Japanese Society for Artitcial Intelligence, 2014*, 2014.

[4] H. Yanagimoto, "Improvement of Sentiment Dictionary Using Neural Network Language Model", *The 28th Annual Conference of the Japanese Society for Artifcial Intelligence, 2014*, 2014

[5] T.Ito, K.Izumi, K.Tsubouchi, T.Yamashita, "Polarity propagation of financial terms for market trend analyses using news articles", *Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN)*, 2016, (To be published)

[6] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient estimation of word representations in vector space", *Proceedings of Workshop at ICLR, 2013*, 2013

[7] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, "Distributed Representations of Words and

Phrases and their Compositionality", *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pp. 3111-3119, 2013.

[8] Y. Yuan, L. He, L. Peng, Z. Huang, "A New Study Based on Word2vec and Cluster for Document Categorization", *Journal of Computational Information Systems 10: 21 (2014)*, pp. 9301-9308, 2014.

[9] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overftting', *Journal of Machine Learning Research 15 (2014)*, pp. 1929-1958, 2014

[10] T. Kudo, K. Yamamoto, Y. Matsumoto, "Applying Conditional Random Fields to Japanese Morphological Analysis", *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pp. 230-237, 2004.

[11] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent Dirichlet Allocation", *Journal of Machine Learning Research - JMLR , vol. 3*, pp. 993-1022, 2003