3L4-OS-26b-1in

株価掲示板データを用いたファイナンス用ポジネガ辞書の生成

Positive / Negative Detection for Finance Contents via Stock Bulletin Boards Data

坪内　孝太 *1　　山下　達雄 *1
Kota Tsubouchi　　Tatsuo Yamashita

*1Yahoo! JAPAN 研究所
Yahoo! JAPAN Research

Judging contents of tweets and blog diary whether positive or negative is useful for successful equity investment. In order for computer to judge such contents automatically, the accurate dictionary specified in finance is important. However, creating the specified dictionary is needed special knowledge for finance, and maintaining the dictionary by manual is so troublesome.

This paper propose the way of creating the dictionary from the stock bulletin board data automatically. Actual test shows that valid dictionary for judging positive and negative is produced.

1. 序論

Twitter やブログなどの投稿を機械が自動的に解析し、投資判断を行う際、記事のポジネガを判定する辞書が必要になる [Wu 11]。一般に辞書の精度は今後のツールや結果の精度に大きく影響する。ポジネガ辞書から記事のポジティブ/ネガティブの度合いをスコア化し、それをベースに学習や相関をとる事が多いが [Celikyilmaz 10]、そもそもの辞書のポジティブ/ネガティブの判断に誤りがある場合、そこから算出されるスコアも無意味なものになってしまうからである。

一般的なポジティブ/ネガティブ辞書であれば、先行研究は多くあり、いろいろと出回っている。しかし、たとえば本稿で対象とする投資関係の場合、一般的な辞書では不十分である。また、企業に関するような評判は、時間軸による影響も多いため、一度作った辞書の見直しは高頻度に求められる。

投資判断に特化した専門的な辞書の作成や維持管理は専門家の知識が必要であり、作成やメンテナンスに大きな労力を要する作業といえる。

そこで、本稿では Yahoo! 株価掲示板におけるコメントデータを用い、ファイナンスに特化したポジネガ辞書を半自動的に生成する手法について述べる。本手法により、特に株式や企業評判に専門知識がない管理者でも専門辞書を生成、管理することができるようになる。

2. 株価掲示板情報を用いた辞書生成手法

2.1 Yahoo! 株価掲示板

本研究では Yahoo! 株価掲示板情報を対象とする。Yahoo! 株価掲示板は、以下の情報からなる。

- 対象銘柄：何の銘柄についての情報か
- タグ情報：強く買いたい、買いたい、様子見、売りたい、強く売りたいの５種
- タイトル：コメントのタイトル
- フリーコメント：自由記入のコメント
- 投稿日時：記事を投稿した日時

連絡先: 坪内　孝太，ヤフー株式会社 Yahoo! JAPAN 研究所，東京都港区赤坂 9-7-1 ミッドタウンタワー，03-6864-3412，ktsubouc@yahoo-corp.jp

当該掲示板は、１）対象銘柄についての議論がなされている事、２）対象銘柄に対する投資意向を示すタグがセットで投稿されている点の２点で他の一般的な掲示板とは異なっているといえる。

2.2 本手法の概要

まず、フリーコメントとタイトルに対し形態素解析を行い、すべての形態素 N-gram を表層文字列としてカウントし、頻度 2 以上の文字列のみを残す。各文字列は「他の文字列 A の部分文字列であり、かつ、A の頻度を超えない」場合、削除される。これにより、頻度 2 以上の最大長の部分文字列が残る。これらをフレーズと呼び、機械学習の素性として使用する。

次に「強く買いたい、買いたい、様子見、売りたい、強く売りたい」の５種のタグをポジティブ/ネガティブに分類する。「強く買いたい」および「買いたい」の文書をポジティブ（+1）に、「売りたい」「強く売りたい」のタグのついた文書をネガティブ（-1）に設定する。

これらの（ポジ/ネガ：フレーズ）のペアを入力とし、フレーズのポジティブ/ネガティブをうまく分類する L2 正則化項を持つ線形回帰モデルを求める。導出されたモデルの各フレーズに設定される重みの正負を調べる事で、各フレーズのポジネガおよびその強度を判定する基準となる。

3. 実証実験

3.1 実験の概要

提案手法の有用性を試す目的で実証実験を行った。

実験には、2012 年 11 月から 2013 年 11 月末までの Yahoo! 株価掲示板情報を用いた。

まず、2013 年 10 月末までの全 3,284,544 件の投稿を用い、モデルの学習を行った。全投稿のうち、ポジティブ/ネガティブのタグ情報が付与されている投稿は約 500,000 件であった。

次に、モデルの精度評価を 2013 年 11 月の 1ヶ月間のデータで試した。評価データは、189,532 件の全投稿のうち、ポジティブ/ネガティブのタグが付与されているものは約 21,034 件であった。これらのポジティブ/ネガティブをどの程度の精度で当てる事ができるかを評価する。

評価は、「ポジティブ（＝強く買いたい、買いたい）」と「ネガティブ（＝強く売りたい、売りたい）」のみをあてる 2 値分

類による評価と、これに「ニュートラル（＝様子見）」を加えた 3 値で当てる評価の 2 種類を行う。2 値分類のの問題では、「様子見」というタグのついたデータは学習・テストともに利用しなかった。モデルの精度は Precision と Recall を求め、評価した。

生成された辞書の性能を評価する基準として、一般的なポジネガ辞書を準備し、ポジティブ/ネガティブ当てのモデルの精度を比較する。一般的な手法の比較では、ポジティブ表現 69 語、ネガティブ表現 102 語からなる一般的なポジネガ表現辞書を使用する。これは、Twitter の分類等に使われるような一般的な辞書である。ポジティブな表現例として、「うれしい」「ありがとう」「楽しい」、ネガティブな表現例として「むかつく」「嫌だ」「キモい」といった言葉が並んでいる。各投稿に対してポジ/ネガどちらの表現が多く現れるかを単純に数えて多い方を判定結果とする。

3.2 実験の結果

3.2.1 定性的な評価

本手法により、生成されたポジティブ/ネガティブ辞書の例を表 1 に示す。表には頻度が 1000 以上のものを抽出した。結果、1888 個の単語からなるポジティブ/ネガティブ辞書が生成された。表はそのうちのポジティブ/ネガティブなスコアの大きかった上位 15 件ずつを掲載したものである。生成された表を見ると、「青天井」「買い時」「上場来高値」「売り/買い煽り」「信用買い」のように、株式投資に特化した辞書が生成されていることが分かる。

また、「ｙ－゜゜゜」や「wwww」といったネットサービス上の独特の表現も抽出できることが分かる。

表 1: 辞書の例（ポジ/ネガをそれぞれ 15 件ずつ）

	単語（ポジ）	スコア		単語（ネガ）	スコア
1	ポコ♪	1.203	1	一発	-1.477
2	w@	0.872	2	買い豚	-0.999
3	ｙ－゜゜゜	0.853	3	糞株	-0.951
4	いつ買う	0.778	4	ストップ安	-0.927
5	青天井	0.766	5	暴落	-0.655
6	売り豚	0.749	6	ナイアガラ	-0.634
7	ＢＩＧ	0.733	7	買い煽り	-0.596
8	買い気配	0.687	8	はははは っ	-0.588
9	買いです	0.582	9	投資は自己	-0.573
10	買い時	0.555	10	（＾ｏ＾）	-0.519
11	買いだ	0.543	11	落ちる	-0.514
12	買い場	0.537	12	割高	-0.500
13	買います	0.533	13	wwww	-0.488
14	指紋	0.523	14	失望	-0.474
15	がんばれ	0.505	15	自社株買い	-0.470

3.2.2 定量的な評価

2 値分類（ポジ/ネガ）および 3 値分類（ポジ/中立/ネガ）による評価の例を表 2 および表 3 に示す。表内の数値は実際の正解のラベルに対して、提案手法により生成された辞書によるポジネガ判断での、Precision, Recall および F 値である。

次に一般的な手法との比較結果を表 4 に示す。一般的なポジティブ/ネガティブ辞書と提案手法により生成されたポジティブ/ネガティブ辞書によりラベルのついた評価データを分類し、その性能を比較した。数値は精度 (Acuracy) を表し、分類性能の高さを意味している。結果、一般手法に比べて 2 値分類、3 値分類どちらのケースに置いても大幅な改善が見られた。

表 2: 2 値分類の分類結果（提案手法）

	Precision	Recall	F
Nega	0.6353	0.5295	0.5776
Posi	0.8872	0.9241	0.9052

表 3: 3 値分類の分類結果（提案手法）

	Precision	Recall	F
Neu	0.2508	0.5110	0.3364
Nega	0.6317	0.2766	0.3848
Posi	0.8106	0.6912	0.7461

表 4: 実験結果：手法の比較

	一般辞書	提案手法
2 値分類	0.522	**0.845**
3 値分類	0.213	**0.590**

4. 考察

実験の結果を見ると本手法の有用性が示されていることが読み取れる。まず、一般的な辞書では対応できないような投資判断の材料を示す一般的な語を抽出できた。最も長い N-gram により単語を抽出しているため、特に形態素解析の結果に前処理を加えなくても単語抽出が実現できている。

定量的な評価を見ると、一般的なポジネガ辞書分類の性能よりは大幅に改善できている点、および実際の 2 値分類においてはポジネガをうまく判断できていることが分かる。しかし、約 3200 件の分類ミスがある。これらは、分類に失敗したのか、そもそも正解として投稿者によりつけられたタグが不適切なのかを調査する必要がある。

また、中立（Neutral）な記事に対する分類は precision が極端に低く、3 値分類全体での AUC が低くなっている。ポジネガ判定において中立を同じ手法で判断するのは難しい。

5. 結論

株価掲示板情報を用い、ファイナンスコンテンツ専用のポジティブ/ネガティブ辞書を作成する手法について提案し、実際のデータにより評価した。

評価の結果、一般的な辞書と比べてファイナンスコンテンツに対して明らかに有効なポジネガ判断ができ、本手法の有効性について確認できた。

参考文献

[Wu 11] Wu Ye, et. al. ”Learning sentimental influence in twitter.” Future Computer Sciences and Application (ICFCSA), 2011 International Conference on. IEEE, 2011.

[Celikyilmaz 10] Celikyilmaz Asli, et. al. ”Probabilistic model-based sentiment analysis of twitter messages.” Spoken Language Technology Workshop (SLT), 2010 IEEE.