1J5-OS-13b-3in

株価掲示板情報における煽り情報の検出

The Agitation Detection in Stock Bulletin Board

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

*1Yahoo! JAPAN 研究所
Yahoo! JAPAN Research

A few users submit the agitate information in the stock bulletin board for control the stock price. For instance, agitator submit fake information or information from opposite stance to actual. These information deteriorates reliability of the bulletin board and be a strong noise in data analysis. This paper propose the way of agitation detection of stock bulletin board by analyzing users' article. The simulation test with 2 years bulletin board data shows the proposed method can detect agitation accurately.

1. 序論

株価掲示板に投稿された情報を解析し、煽り投稿を自動検出することを目指す。

株価掲示板情報には株価の参考になる情報や個人投資家の感想など有用な投稿も多いが、ノイズとなる投稿も多い。例えば、一般的にスパムと呼ばれる宣伝や他のユーザに対する罵倒などである。一方、株価掲示板に特有なものとして、個人の考えとは逆の情報を流す事で他人の売買を誘導する「煽り」投稿がある。煽りは、掲示板ユーザを惑わし掲示板自体の信頼性を損なうものであるため、スパム同様に自動判定による排除が求められる。

本稿では、まず株価掲示板における煽りについての定義し、それに基づき実際のデータから煽り投稿を抽出する実験を行い、抽出の際に使用する手法と精度の関係について論じる。

2. 煽り情報の検出

2.1 対象データ

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 投稿のタイトル

テキスト 投稿の本文

投稿日時 記事を投稿した日時

対象銘柄についての個別の議論がなされている事と、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が他の一般的な掲示板とは異なっているといえる [坪内 2014]。

本研究では約 2 年分の投稿データを使用した。

期間 2012-11-21 から 2014-11-17 (727days)

投稿数 14,015,844

以降、5 種のタグ情報「強く買いたい、買いたい、様子見、売りたい、強く売りたい」をそれぞれ「2,1,0,-1,-2」の数値に置き換え「感情スコア」と呼称する。

2.2 煽りの定義

「煽り」とは何かを明確に判別できるレベルで定義するのは難しい。例えば、「この株は絶対上がる」との投稿が、単なる初心者の感想なのか、株価上昇で利益を得ようという煽りなのか、判別は困難である。株関連の知識、市場の状況、投稿するユーザの個性などを総合的に見て判断する必要がある。

そこで本研究では「掲示板の利用者による判断が総合的な知見に基づく判断に近い」との考えから、他の投稿者から「煽りである」と言及された投稿を「煽り投稿」と見なすこととした。この方針で「煽りとその指摘」の組の抽出を高精度で行うことが本研究の目的である。現段階では実際の煽りか否かは問わず、煽りの指摘である否かに着目する。

下記に「煽りとその指摘」の例を 2 つ挙げておく。

タイトル やっと下げトレンドスタートか

テキスト もしさっきドテン買いしてたら終わってたな
今の株価は異常すぎる
何かの拍子で崩れたら一気に暴落しそうなんだがなあ

タイトル Re: やっと下げトレンドスタートか

テキスト ここのホルダーはそんな売り煽りに負けるほど気の小さい株主達じゃないですよ。

2.3 煽り投稿の検出

前節で述べた方針にそって、煽りとその指摘の組の候補を抽出する。まず下記の条件で全投稿から煽り候補を抽出する。

条件 1 ある投稿（元投稿）に対してのリプライ投稿のテキスト中に「煽」という文字があり、元投稿に「煽」の文字がない

この条件により約 5 万件が抽出できた。目視で、これらの投稿とリプライ投稿の組を 100 件ランダムで確認した結果、58 件が「煽りの指摘」に該当した。表 1 に抽出候補数、表 2 に

連絡先: 山下 達雄，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，tayamash@yahoo-corp.jp

煽り抽出精度を示す (表には以下で説明する条件での結果も挙げている)。

この抽出結果はリプライ投稿に文字「煽」(「煽った」「煽り」「煽る」といった表現) が含まれているのみであり、元投稿に対しての煽りの指摘であるか否か正確には分からない。例えば、「煽りに関するメタな話題」や「リプライ先ではない他の投稿への煽り指摘」など「煽りの指摘ではない」ケースもある。

これらのノイズを排除し「煽りの指摘」の抽出精度を高めるため、本研究では感情スコアの差に着目した。抽出結果の印象として、元投稿の感情スコアとリプライ投稿の感情スコアが正反対である場合、煽り指摘の可能性が高くなる傾向があった。そこで、この抽出結果を下記の条件でさらに絞り込んだ。この条件 2 は、元投稿の感情スコアが「強く買いたい」(2) であったときにリプライ投稿の感情スコアが「強く売りたい」(-2) の場合（またはそれぞれが逆)、煽り指摘の可能性が高くなるという考えに基づく。

条件 2 投稿とリプライ投稿の感情スコアの差が 4 である

抽出結果は 571 件となった。ランダムで 100 件を目視で確認した結果、87 件が「煽りの指摘」に該当した。

条件 1+条件 2 による抽出数は条件 1 のみ場合と比べ量は 1%程度だが「煽りの指摘」の抽出精度は高い。

また、条件 2 単体での抽出精度も調べた。条件 2 で全投稿から抽出した組は約 9 千件となった。ランダムで 100 件を目視で確認した結果、18 件が「煽りの指摘」に該当した。

(リプライ投稿を持つ) すべて投稿とそのリプライ投稿の組についても調べた。全投稿中にリプライ投稿を持つ投稿は約 143 万件あり、ランダムで 100 件を目視で確認した結果、「煽りの指摘」に該当するものは 0 件であった。

表に各方法での「煽りの指摘」の抽出精度を示す。

表 1: 煽り候補抽出数

	条件 2	not 条件 2
条件 1	571	49640
not 条件 1	9397	1437653

表 2: 煽り抽出精度

	条件 2	not 条件 2
条件 1	0.87	0.58
not 条件 1	0.18	0

さらに、煽りの判別に言語表現を使うことで精度向上が可能か確認するため機械学習での分類実験を行った。目視で判定した「煽り」データ 234 件とランダムで選んだ「煽りでない」データ 10000 件に対する SVM による 2 値文書分類タスクで、素性は投稿タイトルとテキストに対する極大文字列である [岡野原 2008]。10-fold cross validation による結果を表 3 に示す。煽りに対する Precision は 0.19 と低く、単純な言語表現だけでの検出は難しいことが分かる。

3. 考察

検出のための一連の検討・実験の結果を見ると、「煽りの指摘」を抽出するためには、煽りを表現する言語情報も重要であるが、先言及との感情スコアの差の方がより重要であることが分かる。

表 3: 機械学習による分類結果

	Precision	Recall	F 値
煽り	0.1905	0.0513	0.0808
煽りでない	0.9782	0.9949	0.9865

煽りの指摘は感情的な行為であるとも言え、それが直接反映されている感情スコアが有効な指標となっている。例えば、下記は条件 1 で抽出された候補であるが、煽りについてのメタな話題であるため、煽りの指摘ではない。読めば分かる通り、二人とも「買いたい」という同じ意見であり煽りでないことが明確である。実際にどちらの投稿も感情スコアが同じであり（2=強く買いたい)、条件 2 でうまくフィルタアウトできている。

タイトル うぅぅ

テキスト ずっと悩んでやっと買った途端の下げ続きにたえられず、昨日の寄りで損きりしてしまった自分が悔 しくて寝付けません
〇〇〇〇ってホントにすごいんですねっ
明日買えるかな … 買いたいな ····

タイトル Re: うぅぅ

テキスト 分割前 150 は堅いと予想します。（オルフェーブルの単勝より堅いかも ?! ）
今日、そして分割までどうなるかわかりませんが、今日は様子を見てから買うと良いかもしれません ね。
売り煽りになりますが、後場でナイアガラが来てほしい (笑)

買いたい・売りたいという行動に結びつく感情の方向が、煽りであるかメタな話題であるかの判別の大きなヒントになっている。

4. 結論

株価掲示板情報を用い、「煽り」を検出するための手法を提案し、実際のデータにより評価した。

評価の結果、言語情報に加えて感情スコアの差という株価掲示板ならではの情報を用いる手法が有効であることが確認できた。

今後の課題としては、実際の株価の動きとの関係や、各ユーザの投稿活動の分析などがあげられる。

参考文献

[坪内 2014] 坪内孝太, 山下達雄: 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, 人工知能学会 全国大会 2014, May 2014.

[岡野原 2008] 岡野原大輔, 辻井潤一: 全ての部分文字列を考慮した文書分類, 情報処理学会研究会報告 NL(187), September 2008. Septembar 2008.