6W-01

SNS を用いた株価の騰落予測における ツイート抽出方法の比較検証

安田健士郎　　櫻井義尚
明治大学　総合数理学部

1　はじめに

近年，機械学習技術が大きく進化し，金融や証券の分野においても機械学習を利用した市場予測が活発に行われるようになった．

最近では，SNS のテキストデータを感情データとして変換して利用し，今後の金融市場を予測する手法が提案されている．SNS は世間的なトレンドや感情がリアルタイムに検知することが可能な情報源として注目されており，金融分野においても有益なデータとして研究に利用されるようになった．

しかし，これまでの研究ではツイートなどのテキストデータをどのように定量化するかという問題に焦点が置かれ，収集されたツイートをどう抽出するべきかについては，詳しく検証されていなかった．

そこで本研究では，ツイートを用いた日経平均株価予測において，有効なツイート抽出手法を明らかにする．具体的には，「日経平均」というキーワードと複数の経済単語を使用したツイート抽出手法を提案し，これらのツイートをそれぞれ感情値に変換した後，時系列モデルに組み込んで騰落予測を行い，その予測精度を比較検証する．

2　関連研究

Bollen ら[1]は，Twitter から得られる集団的な気分状態を 6 種類で定義し，これによって得られる測定値がダウ平均株価の予測に貢献するかを検証した．

迫村ら[2]は，Twitter のツイートデータから得られるテキスト特徴量とグラフ特徴量が経済動向と関連性があるかを明らかにし，これらが経済指標の予測に貢献するかどうかを検証した．

このように，ツイートを特徴量としてどう定量化させるかに焦点が置かれた論文は多いが，ツイートを抽出するための日本語のキーワード選定手法については詳しく検証されていない．

3　提案手法

日経平均株価の騰落予測モデルに使用するツイートデータの抽出手法について提案する．

3.1　Twitter API によるツイートの取得

Twitter API は，ある任意の単語を入力すると，その単語を含むツイートを自動取得する仕組みとなっている．本実験では，日経平均株価の騰落を予測して検証するため，「日経平均」という単語を含むツイートを取得した．

それに加えて，経済に対する Twitter ユーザーの感情を組み込むため，経済関連の単語を含むツイートを取得した．なお，Twitter API にはレート制限が存在するため，経済関連単語を限定する必要がある．よって，本研究では日経シソーラスに収録されている経済関連単語から選定を行った．　具体的には，日経シソーラスの「経済・産業」分野内の「証券」分野に属している上位単語に該当する単語を含むツイートを取得した．

3.2　予測に使用するツイートの抽出手法

「日経平均」と経済関連単語の両方の単語が含まれるツイートを用いることで，予測精度が高くなるという仮説のもと，　3.1 にて取得したツイートを「日経平均」のみを含むツイートデータセット，経済関連単語のみを含むツイートデータセット，「日経平均」あるいは経済関連単語のどちらかが含まれるツイートデータセット，「日経平均」と経済関連単語の両方が含まれるツイートデータセットの４つの手法に分けて抽出した．

Comparison of tweet extraction methods for predicting stock price rises and falls
†Kenshiro Yasuda　†Yoshitaka Sakurai
†Meiji University

4　実験

4.1　実験データ

日経平均株価のデータは 2017 年 9 月から 2018 年 2 月の期間で YAHOO!JAPAN ファイナンスより取得した．なお，土日祝日に関しては株式市場が閉鎖されているため，その日は欠損値とする．今回は，騰落を予測するため，前日比との変動差で減少した場合は「0」，変動無し或いは増加した場合には「1」の騰落ラベルを付与させた．

ツイートデータは株価と同様の期間にて，3.1 の手法を用いて取得した．その後，MeCab を用いてそれぞれのツイートを形態素解析し，前処理として URL や記号をツイートから削除した．最後に，前処理を施したツイートを高村ら[3]の作成した単語感情極性対応表辞書を利用して感情値化し，日付毎に平均化させた．

4.2　予測モデルの構築

本実験では時系列情報を考慮するため，LSTM を用いた深層学習モデルを構築する．このモデルの構造は以下の図 1 の通りで，「騰落ラベル」と「ツイートの感情値」を入力層に入力し，それぞれを LSTM 層に経由させて連結層で結合させる．結合後は全連結層を通過し，不活性化率 25%の Dropout 層を経由して出力層にて出力させる．騰落の 2 値分類予測のため，出力層では Sigmoid 関数を用いて出力値を 0 から 1 の間に収めている．

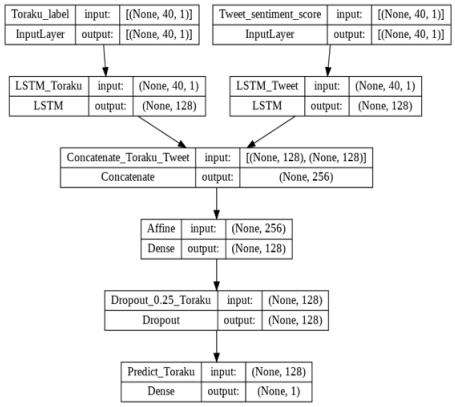

図 1　予測モデルの構造

4.3　検証実験

ツイートの感情値のデータセットをそれぞれ株価の騰落ラベルと同時に予測モデルに入力し，騰落予測を行う．入力した全てのデータ 120 日分のうち，学習データを 96 日，テストデータを 24 日に振り分けて予測の精度を測り，3.2 で述べた 4 つの抽出手法の結果を比較する．

4.4　予測結果と考察

表 1 に騰落予測実験の結果を示す．なお，評価指標は分類モデルのため，Accuracy，Precision，Recall，F-measure を採用した．

表 1　実験結果

	日経平均のみ	日経シソーラス上位単語のみ	日経平均 OR 日経シソーラス上位単語	日経平均 AND 日経シソーラス上位単語
Accuracy	0.483	0.375	0.458	**0.533**
Precision	0.411	0.351	0.397	**0.455**
Recall	0.580	0.640	0.560	**0.660**
F-measure	0.478	0.449	0.462	**0.534**

日経平均と日経シソーラスの上位単語の両方の単語が含まれるツイートの抽出手法が 4 つのどの指標においても一番予測精度が高かった．

この結果から，日経平均と経済関連の単語を用いて日経平均株価について言及しているツイートを抽出することで，経済トピックに関する具体性が増し，Twitter ユーザーによる世間的な感情として日経平均株価に反映されやすいのではないかと考えられる．

5　まとめ

本研究では，ツイートの感情値を用いた株価の騰落予測モデルに使用するためのツイートの抽出手法について提案し，比較検証を行った．

実験結果から，「日経平均」と経済関連単語の両方を含むツイートを用いる抽出手法が，他の抽出手法より予測精度が高かった．

なお，分類モデルの評価指標としては全体的に精度が低く，各指標においても他の抽出手法と大幅な差異が見られないものもあった．よって，モデルの構築やツイートの感情値化の部分で改善が必要とされる点が今後の課題として挙げられる．

参考文献

[1] Bollen, J., Mao, H. and Zeng, X. “Twitter mood predicts the stock market”. J.computational Science, Vol.2, No.1, pp.1–8, 2011

[2] 迫村光秋，和泉潔，セーヨー・サンティ．“Twitter のテキストとネットワークの解析による経済動向分析”．第 10 回人工知能学会研究資料

[3] 高村大也，乾孝司，奥村学，″スピンモデルによる単語の感情極性抽出″，情報処理学会論文誌ジャーナル，Vol.47 No.02 pp. 627--637，(2006)