研究指導 中澤 真 教授

決算短信に用いられる単語の出現傾向と株価変動の関係性分析

田部 望

1. はじめに

近年，さまざまな分野でテキストマイニングを活用した研究が行われている[1]．その一つとして，金融分野でのテキストマイニングを用いた市場予測や企業分析があげられる[2]．これまでは株価や経済指標など定量データでの分析が中心だったが，技術の進歩によって大量のテキスト情報の分析も比較的容易にできるようになった．テキストの情報源としては，Twitterや掲示板，ニュース記事，経済リポートなど書き手や文章量などを問わず多岐にわたっている．

しかし，テキスト情報は指標化が難しく，分析の際には分析者の経験や勘に頼る部分が大きくなるため，客観的な判断が難しい．有価証券報告書を分析するWebサービスの「有報リーダー」[1]なども存在するが，単語の出現頻度を算出するような機能はないため，テキストの定量的な分析はできない．金融分野でテキストマイニングを活用する取り組みも行われているが[3][4][5]，有用な情報の抽出を試みたものが主であり，単語の出現頻度の傾向と株価との関係を分析するような試みはまだ行われていない．

そこで，本研究では企業が発表する決算文書を対象にテキストマイニングを用いた定量的な分析をし，出現する単語と株価の関係性を明らかにすることを目指す．

2. 決算文書のテキストマイニング

2.1 金融分野におけるテキストマイニングの概要

投資判断や企業分析には，決算文書やニュース記事などのテキスト情報を利用することが多い．しかし，これらの文書は文章量が多く企業数も膨大であるため，人手によって分析するのは非効率である．そこで，テキストマイニングを用いて効率的に文書の分析を行う研究がなされている[3][4][5]．これによって，定量データだけではできなかった有用な文の抽出などが実現できるようになった．

2.2 企業の決算文書

本研究では，企業が発表する重要な情報である決算文書に注目する．決算文書の代表的なものとして，有価証券報告書と決算短信の2つがあげられる．有価証券報告書は金融商品取引法によって提出が義務付けられた文書であり，情報量が多く信頼性も高いが決算後3か月以内の発表であるため速報性は低い[6]．一方，決算短信は正式な決算発表ではないが，決算後30日から45日ほどで発表されるため有価証券報告書と比較して速報性が高い．いずれの文書も財務諸表などの定量データ以外に企業の概要や経営成績などといった情報がテキストとして多く記載されており，企業の評価や投資判断の材料として用いられている[7]．このように，決算文書は企業の業績などを網羅的に知ることができる重要な情報であるため，本研究では決算文書を分析し，出現する単語と株価との関係性を明らかにする．

2.3 決算文書分析の現状

決算文書をテキストマイニングで分析した研究はいくつかある．竹内ら[3]は有価証券報告書を分析し，倒産企業を特徴づける表現の抽出を試みている．北森ら[4]は決算短信から業績予測文を抽出する手法を提案している．斎藤ら[5]は決算短信中の単語をもとに企業間の関係性の抽出を試みている．このように多くの研究が企業の評価や投資判断の支援をする目的でなされている．

しかし，決算文書をテキストマイニングして株価予測を試みた研究は，筆者が知る限りいまだ存在しない．単語の出現頻度はテキストマイニングに詳しくない人にとっても数値的に判断しやすい指標であるため，得られた知見を多くの人が活用できるようになることが期待される．

そこで，本研究では決算短信のテキスト部分を分析し，出現する単語と決算短信公開後の株価変動との関係性を明らかにすることを目指す．そして，ランダムフォレストを用いて決算短信を株価の変動に応じて分類し，その分類の精度を検証する．これによって分類結果をもとに株価の予測が可能かどうかを判断することができる．なお，決算短信を対象とした理由は，速報性が高いという文書の性格上，株価への影響も大きいと考えられるためである．

3. 決算短信の分析

3.1 分析対象

本研究では決算短信に出現する単語と株価の関係性について，東証一部上場の銀行を対象に分析する．銀行業では企業ごとの業務内容の差が小さいと考えられる．そのため，それぞれの企業固有の商品やサービスに関連する単語の影響をあまり受けず，

[1] http://uforeader.com/v1/

業績や株価にかかわるような単語の抽出が容易になると考えたためである. なお, 調査対象データは2016年3月期と2017年3月期のものとする.

決算短信は各企業のWebページからPDF形式のファイルをダウンロードし, これを「Adobe Acrobat Reader」[2]を用いてテキスト形式に変換した. これにより2016年3月期は72件, 2017年3月期は81件の合計153件の文書を取得した. 株価については「株式投資メモ・株価データベース」[3]から取得した.

分析にあたり, 決算短信テキストに対し全角と半角の統一, 記号や英数字の削除などの前処理を行っている. なお, 分析には統計解析ソフト「R」[4]と形態素解析器「MeCab」[5]を使用した[8].

3.2 株価変動の計算

株価は, 政治や経済など市場全体の情報を表す一般経済情報と企業の決算発表など個別の情報を表す個別企業情報の影響によって変動する. そのため, 一般経済情報の部分を取り除くために市場リターン控除法を用いて計算する[9]. まず決算短信の公開日とその前後10日間(合計21日間)の株価(調整後終値)を用いて個別リターンを算出する. 次にTOPIXを用いて市場リターンを算出し, 個別リターンから差し引くことによって個別企業情報による株価変動部分である異常リターンを算出する. 最後に21日間の異常リターンを累積して累積異常リターンを算出する. これにより, 決算短信の公開によって当該企業が市場平均を上回る, または下回る異常リターンを累積的に獲得したかどうかを表すことができる. この累積異常リターンの基本統計量をまとめた結果を以下に示す.

表 1 累積異常リターンの計算結果

	文書数	中央値	平均値	最大値	最小値
2016年	72件	-1.37%	-1.09%	11.20%	-14.40%
2017年	81件	-5.05%	-5.08%	7.80%	-27.60%

3.3 出現頻度上位の単語の比較

表 1より, 累積異常リターンの分布には十分な幅があり, 株価上昇幅の大きい企業と下落幅の大きい企業があることが確認できた. そこで, 株価の上昇・下落を特徴づける決算短信内の単語の抽出を試みるために, 累積異常リターンが正の値の企業と負の値の企業の2つの群に分け, それぞれの群の単語の出現頻度を求めた. 2つの群で出現頻度の傾向に明確な差が現れれば, これらの単語に基づき株価の予測ができることになる. 2016年3月期の決算短信での結果を表 2に示す. なお, 出現頻度の上位10単語を抽出しており, 品詞は斎藤ら[5]の手法に基づき名詞, 動詞, 形容詞の3種類を対象とした.

この表から出現頻度上位の単語は業種特有の単語や株価の上昇と下落の判別に直結するような単語がないことがわかる. また, すべての単語が上昇と下落の両方の群に出現している. そのため, 出現頻度上位の単語の有無だけで, 株価の上昇・下落のいずれの群に属する決算短信であるかを判別するのは困難であるといえる.

表 2 2016 年 3 月期決算短信の頻出上位 10 単語

単語	品詞	累積異常リターンが正の群での順位	累積異常リターンが負の群での順位
する	動詞	1	1
平成	名詞	2	2
万	名詞	3	4
百	名詞	4	3
年月	名詞	5	9
その他	名詞	6	5
株式	名詞	7	8
年月日	名詞	8	6
連結	名詞	9	7
資本	名詞	10	10

4. ランダムフォレストによる分類

4.1 TF-IDF 値の算出

前節で出現頻度をもとに株価の上昇と下落をそれぞれ特徴づける単語の抽出を試みたが, 特徴的といえる単語の抽出はできなかった. 一般に, 文書の特徴を表す単語を抽出する際には, 単語の重みづけをする. テキストマイニングの分野では, TF-IDFという重みづけの指標がよく用いられる. TF-IDFとはTF(Term Frequency) と IDF(Inverse Document Frequency)という2つの値を組み合わせた指標である[10]. TFとIDFはいずれも数種類の計算方法があるが, 今回はTFとして索引語頻度, 対数化索引語頻度, 2進重みの3種類, IDFとして逆文書頻度とエントロピーの2種類の評価値を用いて算出する. TF-IDF値は長い文書ほど大きな値になる傾向があるため, 文書の長さに応じた正規化も行う必要がある. 今回はコサイン正規化を行い, 文書ごとのTF-IDF値の二乗和を1にする. このような複数の評価尺度によるTF-IDF値によって文書をベクトルとして表現したものを用い, どの尺度に基づき分類すれば精度が高くなるかを検証する.

4.2 ランダムフォレストとは

TF-IDF値によって作成した文書ベクトルを指標として分類を行い, 株価が上昇した企業群と下落した群にどの程度分類できたか, その精度を検証する. 本研究ではデータ量が大きくても効率的に動くという利点からランダムフォレストを分類アルゴリズムとして採用した. ランダムフォレストはBreiman[11]によって

[2] https://acrobat.adobe.com/jp/ja/acrobat/pdf-reader.html
[3] https://kabuoji3.com/stock/
[4] https://www.r-project.org/
[5] http://taku910.github.io/mecab/

2001年に提案された機械学習アルゴリズムである．文書分類の分野では木村[12]などで用いられており，一定の精度を示している．また，ランダムフォレストは分類に寄与した変数を抽出することができる．これによって分類に有効な単語を抽出することで，株価予測のための新たな知見を得ることが期待できる．そこで，本研究でも分類に寄与した単語を明らかにするために変数の抽出を行う．

4.3 ランダムフォレストによる分類実験

4.3.1 2 ラベルでの分類

決算短信を株価が上昇した群と下落した群の2群に分けてランダムフォレストを実行する．2016年3月期の決算短信を学習用データとしてランダムフォレストの分類モデルを作成し，評価用データの2017年3月期の決算短信に対して分類を行い，その精度を検証する．上昇と下落は累積異常リターンが正の値のものを上昇，負の値のものを下落として正解のラベルを付与した．ラベルごとのデータ数を以下に示す．

表 3　上昇・下落の 2 ラベルでのデータ数（件）

	上昇	下落
学習用データ	28	44
評価用データ	13	68

4.1節で述べた複数の評価尺度によるTF-IDF値を指標として分類した結果を表 4に示す．値はモデルの作成と評価用データの分類を10回実行した結果の平均値である．分類精度の評価には適合率を用いた．網羅性の指標である再現率が低くとも，適合率が高ければ株価の上昇する企業，逆に下落する企業を判別することが可能になり，株式売買の判断材料として使うことができるからである．なお，太字で表示した数字は各群で最も高い適合率を表す．結果から，上昇群・下落群ともに適合率が最も高い値を示したのはTFに2進重み，IDFにエントロピーを用いたときであることが読み取れる．その場合，下落群の適合率は83.9%を示したが，上昇群の適合率は15.9%であった．そのため，精度向上に向けたさらなる工夫が必要である．

表 4　2 ラベルでの分類結果（%）

分類の指標		評価用データの分類結果	
TF	IDF	上昇群の適合率	下落群の適合率
索引語頻度	逆文書頻度	12.4	83.5
	エントロピー	9.0	82.8
対数化索引語頻度	逆文書頻度	9.5	83.0
	エントロピー	9.4	82.7
2進重み	逆文書頻度	12.4	83.3
	エントロピー	**15.9**	**83.9**

4.3.2 3 ラベルでの分類

表 4では学習用データにおける上昇群と下落群の構成比が偏っていたため，文書数の多い下落と判断される文書の割合が高くなったと考えられる．そこで，上昇群と下落群で文書数をそろえて分類をすることにした．また，データには累積異常リターンの絶対値が1を下回るような株価の変動が微小であるものも存在している．このようなデータは上昇群・下落群の判別が困難だと考えられるため，新たに中間のラベルを設けそこに分類されたものは判別の対象から除外することとした．これによって株価の変動が大きい文書の分類精度向上を目指す．まず，学習用データのうち累積異常リターンの上位3分の1を上昇群，下位3分の1を下落群，残りの3分の1を中間群として文書数を均等にした．なお，評価用データのラベルは表 3のまま変えていない．これは，表 3に示したとおり累積異常リターンの正負によって2ラベルに分けた時点で上昇群と下落群のデータ数に偏りがあったため，さらにラベルを増やすことでデータ数の偏りが大きくなることを防ぐためである．この時の分類結果を表 5に示す．データ数が不均衡だった表 4での結果と比較してわずかながら精度向上がみられ，データ数の均衡化が効果を発揮したと考えられる．

表 5　3 ラベルでの分類結果（%）

分類の指標		評価用データの分類結果	
TF	IDF	上昇群の適合率	下落群の適合率
索引語頻度	逆文書頻度	14.5	80.1
	エントロピー	13.8	81.0
対数化索引語頻度	逆文書頻度	14.7	83.9
	エントロピー	**18.4**	**84.5**
2進重み	逆文書頻度	15.3	82.2
	エントロピー	16.0	82.1

学習用データにおける累積異常リターンの分布は表 1で示したように幅広いものになっていた．表 5の実験では学習用データを単純に3分割したため，各群での累積異常リターンの分布が広くなっていたことが考えられる．株価予測では変動の大きい企業がわかることが重要である．そのため，より株価の変動が大きい文書の分類精度を向上させるために上昇群と下落群の学習用データの数を変化させる．上昇群と下落群は累積異常リターンの上位と下位からそれぞれ10%，20%，30%と抽出する割合を変化させて分類を行った．これは，株価の変動が大きい文書の割合を変化させた際に，精度にどのような影響をもたらすかを確認するためである．中間群は上昇・下落と同数になるようにランダムに抽出し，文書数をそろえた．学習用データの数は10%の場合は各7件，20%の場合は各14件，30%の場合は各22件となった．

各割合での分類の結果を表 6から表 8に示す．2ラベルでの分類と比較して，10%の場合ではどのTF-IDF値でも上昇群の適合率に大きな変化は見られなかったものの，下落群の適合率が最高で94.5%

と高い値になった．20%の場合では上昇群・下落群ともに適合率が低下した．30%の場合は下落群の適合率がわずかに低下したが上昇群の適合率は向上した．全体的な傾向として，TFに対数化索引語頻度か2進重みを，IDFにエントロピーを用いたときに他のTF-IDF値を用いたときと比べて両群とも適合率が上がることがうかがえる．これは表 2に示したような出現頻度上位の単語の影響を小さくできたことが要因として考えられる．このことから，決算短信の分類に際してはTFに対数化索引語頻度か2進重みを，IDFにエントロピーを用いることが有用であるとわかった．しかし，上昇群の適合率は最も高いものでも20.5%であり，実用的な精度とはいえない結果になった．

表 6 抽出割合 10%の場合の分類結果(%)

分類の指標		評価用データの分類結果	
TF	IDF	上昇群の適合率	下落群の適合率
索引語頻度	逆文書頻度	8.3	90.9
	エントロピー	10.0	91.6
対数化索引語頻度	逆文書頻度	12.5	92.8
	エントロピー	12.0	92.1
2進重み	逆文書頻度	9.4	92.3
	エントロピー	**15.4**	**94.5**

表 7 抽出割合 20%の場合の分類結果(%)

分類の指標		評価用データの分類結果	
TF	IDF	上昇群の適合率	下落群の適合率
索引語頻度	逆文書頻度	1.5	73.8
	エントロピー	5.2	74.3
対数化索引語頻度	逆文書頻度	3.1	**77.4**
	エントロピー	5.8	75.7
2進重み	逆文書頻度	**10.3**	75.9
	エントロピー	10.1	75.3

表 8 抽出割合 30%の場合の分類結果(%)

分類の指標		評価用データの分類結果	
TF	IDF	上昇群の適合率	下落群の適合率
索引語頻度	逆文書頻度	15.0	79.6
	エントロピー	15.4	79.6
対数化索引語頻度	逆文書頻度	18.6	82.2
	エントロピー	**20.5**	**84.2**
2進重み	逆文書頻度	14.5	80.5
	エントロピー	14.4	80.5

4.4 分類に有効な変数の抽出

適合率が高い値を示した学習用データの抽出割合が30%の場合でTFとして対数化索引語頻度，IDFとしてエントロピーを用いたときの分類に有効な変数を抽出した．結果の上位5変数を以下に示す．

- 取り組む
- 信頼
- 目次
- 繰延
- 全部

いずれもどの文書にも出現するような単語であるが，TF-IDF値を指標とした際には分類に有効であることが示された．このような単語に注目して分析することで，株価変動の予測精度を向上できると考えられる．

5. むすび

本研究では決算短信のテキスト部分を分析してTF-IDF値を算出し，それを指標としてランダムフォレストを実行し株価に応じた分類を試みた．結果として，TF-IDF値のうちTFに対数化索引語頻度か2進重み，IDFにエントロピーを用いることで最も高い適合率は株価の上昇群では20.5%，下落群では94.5%を示し，下落群は高い予測精度で分類できることがわかった．これにより，株式の空売りをする際の予測の一助とすることが期待できる．

今回は決算短信を用いたが，文章量が少ないため文書の特徴を表す単語の抽出が難しくなってしまい，上昇群と下落群で適合率に差が生じた可能性が考えられる．そのため，有価証券報告書などの文章量が多い文書を用いて分析し，精度を検証することが今後の課題としてあげられる．

参考文献

[1] 齋藤朗宏，“日本におけるテキストマイニングの応用”，The Society for Economic Studies The University of Kitakyushu Working Paper Series No. 2011-12，2012

[2] 和泉潔，松井藤五郎，“金融テキストマイニング研究の紹介”，情報処理 Vol.53 No.9，pp.932-937，2012

[3] 竹内広宣，荻野紫穂，渡辺日出雄，白田佳子，“テキストマイニングによる倒産企業分析”，経営情報学会全国研究発表大会要旨集，pp.124-127，2008

[4] 北森詩織，酒井浩之，坂地泰紀，“決算短信 PDF からの業績予測文の抽出”，電子情報通信学会論文誌 D Vol.J100-D No.2，pp.150-161，2017

[5] 斎藤祐一郎，西森丈俊，“自然言語処理を用いた企業相関関係の取得”，情報処理学会研究報告 Vol.2010-IOT-11 No.4，pp.1-5，2010

[6] 三井住友アセットマネジメント，わかりやすい用語集 解説［有価証券報告書］，https://www.smam-jp.com/glossary/YST1785.html

[7] SMBC 日興証券，決算短信｜初めてでもわかりやすい用語集，https://www.smbcnikko.co.jp/terms/japan/ke/J0041.html

[8] 石田基広，“R によるテキストマイニング入門 第 2 版”，森北出版，2017

[9] 須田一幸，山本達司，乙政正太 編著，“会計操作―その実態と識別法、株価への影響”，ダイヤモンド社，2007

[10] 北研二，津田和彦，獅々堀正幹，“情報検索アルゴリズム”，共立出版，2002

[11] Leo Breiman，“Random Forests”，Machine Learning Volume 45 Issue 1，pp 5-32，2001

[12] 木村美紀，“TF-IDF を用いた文書分類の試み”，文学研究論集第 48 号，pp.1-17，2018