適時開示情報の業績に対するリスク有無の自動判定

Automatic classification according to risk to the corporate performance contained in timely disclosures of company information

矢野 大輔[1]　酒井 浩之[1]　北島 良三[1]
広井 康男[2] 村山 勉[2] 河合 継[3] 西山 昇[4]
Daisuke Yano[1], Hiroyuki Sakai[1], Ryozo Kitajima[1]
Yasuo Hiroi[2], Tsutomu Murayama[2], Kei Kawai[3], Noboru Nishiyama[4]

[1]成蹊大学
[1]Seikei University
[2]株式会社 QUICK
[2] QUICK Corp.
[3]クリスタルメソッド株式会社
[3] Crystal method co. ltd
[4]Dragons' Desk Limited / 千葉商科大学
[4]Dragons' Desk Limited / Chiba University of Commerce

Abstract:
In this paper, we propose a method for classifying timely disclosures of company information with risk to the corporate performance by using deep learning. For example, our method classifies a timely disclosure "A notice of recording a extraordinary loss" as "Including risk". Moreover, our method classifies the timely disclosure as "Extraordinary loss". Our method makes timely disclosures of company information easy to see and investors will be useful to make investment decisions. We evaluate our method and it attained 87.4% accuracy.

１．はじめに

近年，個人投資家の数が増加している中，適時開示情報への関心が高まっている．適時開示情報とは上場企業が義務付けられている「重要な会社情報の開示」のことであり，適時開示情報の中には，その上場企業の株価に影響を与える可能性の情報もある．特に，例えば「業績予想の修正」や「営業停止処分」といった業績にリスクのある情報は株価への影響も大きく，投資判断にも大きな影響を与える．

しかし，適時開示情報はインターネット上でいつでも閲覧できるが[1]，常に新しい情報が掲載され続けているため，すべてを閲覧することは困難であり，さらに，その中から業績にリスクのある情報のみを判別して閲覧することは多大な労力を必要とする．

そこで，本研究では，上場企業が公開する適時開示情報を，深層学習によって業績にリスクがあると考えられる情報のみを自動で抽出し，それらを分類する手法を提案する．例えば，「業績予想の修正に関するお知らせ」の適時開示情報を「リスクあり」と判定し、「特別損失」に分類する．

本研究により，業績にリスクのある適時開示情報の閲覧を容易にし，投資家の投資判断に役立てることを目的とする．

関連研究として，例えば，企業の発行している「決算短信」をテキストマイニングの技術を用いて解析し，経済市場を分析する研究などが行われている[1][2][3][4][5]．酒井らは企業の決算短信 PDF から業績要因文を自動抽出する研究を行っている[1]が，業績要因のみでは，業績に対するリスクを判断することはできない．それに対して，本研究ではリスクの有無を判定できる点が異なる．

[1] https://www.jpx.co.jp/listing/disclosure/index.html

２．提案手法

2.1. 手法概要

Step1: 適時開示情報の中から業績にリスクがあると考えられる情報を人手で抽出して学習データとし，その学習データを後述のワードリスト（表１）に示す語により分類する

Step2: 学習データ・テストデータを適時開示情報の文書ごとに Doc2Vec によりベクトル化する

Step3: 深層学習におけるモデルの最適な中間層や batch を決定する

Step4: Chainer を用いて学習データを用いて学習を行い，Chainer のモデルを作成する

Step5: 作成されたモデルに基づき，リスクがある文書をリスクあり，リスクなしに分類し，さらに，リスクありと判定された文書を，その内容に基づいて分類する．

2.2. 使用するデータ

学習データには 2017 年の適時開示情報を使用し，テストデータには 2016 年の適時開示情報を使用する．

学習データの作成方法を以下に述べる．表１に示したワードリストにある語が含まれている文書を「リスクあり」とし，含まれていない単語を「リスクなし」とする[2]．次に「リスクあり」とした文書の中から「特別損失」,「違反」,「その他」の 3 種類に分類する．「その他」には「火災」,「訴訟」,「損害」といった情報が含まれている．分類には表２に示されている語とラベル名に基づいて分類している．分類した学習データの構成を表４に示す．

表１　ワードリスト（一部）

災害，紛争，テロ，地震，風水害，疫病，パンデミック，国際紛争，訴訟，法改正，知的財産侵害，事件，事故，不正，金融犯罪，コンダクトリスク，（以下略　全９９個）

表２　ラベル名の設定

ラベル名	含まれている語（一部）
特別損失	特別損失，減損損失
違反	違反，不正
その他	災害，紛争，訴訟
リスクなし	/

表３　学習データの構成

ラベル名	文書数
特別損失	164
違反	85
その他	73
負例	453
合計	775

以下に，各ラベルに付与された，学習データにおける文書の表題の例を示す．

特別損失	特別損失（有価証券評価損）の計上に関するお知らせ
	固定資産の減損処理に伴う特別損失の計上に関するお知らせ
違反	建設業法に基づく営業停止処分について
その他	当社連結子会社の火災事故に関するお知らせ
	米国外たばこ事業の買収完了について
	連結業績予想の修正に関するお知らせ
	当社に対する訴訟の提起に関するお知らせ
リスクなし	連結子会社における販売用不動産の売却に関するお知らせ
	株式分割による１株に満たない端数処理にともなう自己株式の買い取りに関するお知らせ

[2] 表１のワードリストは適時開示情報からのリスク情報の判定に経験のある人が作成した．

表４　テストデータの構成

ラベル名	文書数
特別損失	165
違反	85
その他	75
リスクなし	454
合計	779

2.3. 深層学習に使用するモデル

Chainer によって学習データからモデルを作成する．中間層を決定する上で適したモデルを作成するために，次のような手順を考える．モデルには、

「入力層 → X → X → X → 出力層」

となる多層パーセプトロンを用いる．

Step1: epoch を 30 として中間層 X のユニット数を変化させて，テストデータにおける精度を比較する

Step2: 最適な中間層 X におけるユニット数を中間層に使用する

Step3: batch の値を変化させて，テストデータにおける精度を比較する

Step4: 最適な batch の値を使用する

表５　X の値ごとの精度比較

X	epoch	学習データにおける精度	テストデータにおける精度
10	100	0.92	0.83
100	100	0.99	0.88
200	100	0.99	0.88
500	100	0.99	0.88
800	100	1.00	0.89
1000	100	1.00	0.88

表６　batch ごとの精度比較

batch	epoch	学習データにおける精度	テストデータにおける精度
10	30	1.00	0.89
50	151	1.00	0.90
100	200	0.99	0.88
200	200	0.97	0.89
400	400	0.97	0.88
600	600	0.96	0.89

表５・表６によると X や batch の変化によって精度に大きな違いは見られなかった．そのため使用する中間層 X のユニット数は，過学習を避けるため最も早く「学習データにおける精度」が 1 になった 800 を使用し，batch には同様の理由で 50 を使用することにした．

入力層は Doc2Vec によって Wikipedia から約 500MB の記事を次元数 400 で学習させたモデルに基づき，学習データの文書をベクトル化したベクトルを使用する．学習データではなく Wikipedia を使用した理由は，学習データのみでは Doc2Vec の学習に必要な十分なデータがなかったためである．出力層は（２．２）により分類ラベル数の 4 とした．活性化関数には ReLU 関数を使用した．

表７　活性化関数比較

活性化関数	学習データにおける精度
ReLU	0.90
tanh	0.88
sigmoid	0.65

３．評価

本手法の評価を行った．評価用の正解データは，テストデータとした 2016 年の適時開示情報から，表１で示したワードリストに基づき作成した．表４にテストデータの構成を示す．

分類の評価結果を表８・表９に示す．分類の全体精度は 87.4%であった．

表８　分類結果の精度

ラベル	精度
特別損失	95. 2%
違反	88. 2%
その他	74. 7%
リスクなし	89. 4%
全体	87. 4%

最も低い精度であったラベルは「その他」で 74.7%，最も高い精度であったラベルは「特別損失」で 95.2%であった．これは，「特別損失」の文書の特徴は掴めているが，「その他」は特徴が掴みづらかったと言える．

以下に，本手法によって「リスクあり」と判定された適時開示情報の例を表９に示す．

表９　「リスクあり」と判定された
適時開示情報の例

	抽出された記事の例
特別損失	特別損失の計上に関するお知らせ当社は、平成 29 年 9 月期第 1 四半期決算において、特別損失を計上する必要が生じましたので、お知らせいたします。
違反	公正取引委員会からの排除措置命令について本日、当社は公正取引委員会から防衛装備庁が発注する特定ビニロン製品の入札に関して独占禁止法に違反する行為があったとして、下記のとおり排除措置命令を受けました。
その他	インドネシア子会社の火災発生に関するお知らせ 当社の連結子会社であるPT.TOTOKUINDONESIA に隣接する他社工場で火災が発生し、その影響でトウトクインドネシアの工場が類焼しました。
リスクなし	組織改定ならびに執行役の管掌変更に関するお知らせ当社は、平成 29 年 1 月 31 日開催の取締役会において、下記のとおり組織改定ならびに執行役の管掌変更を行いましたのでお知らせいたします。

４．考察

表８より,各ラベルとも精度が 80%を超えており,良好な精度を達成しているが,「その他」における精度が 74.7%と低い結果となっている．これは「その他」には「火災」や「損害」に関する文書など様々な文書が含まれているため，区分の幅が広くなってしまったためであると考えられる．

一方で,「リスクなし」の精度は 89.4%と高い精度が得られており,「リスクあり」と「リスクなし」の分類は高い精度で分類できているということが言える．

本手法により,ワードリストで設定した語が含まれていないにもかかわらず，正しく「リスクあり」と判定された例があった．以下に例を示す．

・「リスクあり」と判定した例

（略）損益の状況<連結決算の概況>平成 29 年 3 月期第 3 四半期決算総括 1 実質業務純益<1>は、連結子会社からの利益寄与が増加した一方、単体の資金関連利益の減少等により、前年同期比 205 億円減益（略）

これは表 1 のワードリストに載っているワードは含まれていないが，類似する単語を抽出して分類できたものであると考えられる．

・「リスクなし」と判定した例

平成 28 年 1 月 5 日にノースカロライナ州ナッシュ郡ロッキーマウントのゲートウェイブルバード 200 に位置するホテルで発生した火災に関して、（略）　今後の見通し本件和解金は損害保険により支払われるため、当社には財務上の負担はなく、平成 30 年 3 月期連結業績への影響はありません。

これは「火災」という単語が含まれているが，負例と判定している．

以上のようにワードリストの単語による抽出では得られないような文書を,「リスクあり」の文書の特徴を学習することによって正しく分類ができているということが分かった．

５．むすび

本研究では,深層学習によって適時開示情報を「特別損失」「違反」「その他」「リスクなし」の４種類に分類する手法を提案した．評価の結果として全体精度は 87.4%であったが，ワードリストの単語の有無に関わらずに文書の特徴を学習することができていたことが分かった．また本研究の分類器を実際の業務に適用することを考える際に,「リスクあり」である適時開示情報を「リスクなし」と判断してしまったり,「リスクなし」であるものを「リスクあり」と判断してしまうことは大きな問題である．そのためリスクの有無の分類が非常に重要な点であるが，本研究では「リスクなし」における精度が 89.4%と高い精度を得ることができた．

本研究では，適時開示情報からリスクの有無を自動判別する手法の提案であったが、実務への応用を考慮すると、判別されたリスクが当該企業以外に派生するリスクを同時に把握することが求められると考える．つまり、企業活動に重要な影響を与える内容を含んだ発表文書の内容に応じ、さらに当該発表

企業と取引関係、資本関係、競業関係等、影響が及ぶと思われる関連先企業をその関係性を含め抽出することができれば実務への応用が期待できる.

参考文献

[1] 酒井浩之，西沢裕子，松並祥吾，坂地泰紀,“企業の決算短信 PDF からの業績要因の抽出”,人工知能学会論文誌, vol.30, no.1, pp.172-182, 2015.

[2] 坂地泰紀，酒井浩之，増山繁，“決算短信 PDF からの原因・結果表現の抽出”，電子情報通信学会論文誌 D, vol.J98-D, no.5, pp.811-822, 2015.

[3] Shiori Kitamori, Hiroyuki Sakai, Hiroki Sakaji, “Extraction of sentences concerning business performance forecast and economic forecast from summaries of financial statements by deep learning”, IEEE Symposium on Computational Intelligence for Financial Engineering & Economics (IEEE CIFEr'17), Hawaii, November, 2017.

[4] 酒井浩之，松下和暉，“決算短信からの業績要因文の抽出”，第 11 回テキストアナリティクス・シンポジウム, pp.87-91, 2017.

[5] Hiroki Sakaji, Risa Murono, Hiroyuki Sakai, Jason Bennett, Kiyoshi Izumi, “Discovery of Rare Causal Knowledge from Financial Statement Summaries”, IEEE Symposium on Computational Intelligence for Financial Engineering & Economics (IEEE CIFEr'17), Hawaii, November, 2017.

[6] TDnet 適時開示情報閲覧サービス， 2018 アクセス https://www.release.tdnet.info/inbs/I_main_00.html