有価証券報告書を活用したテキスト分類モデル作成および業績予測アプリの開発

星野 磨寿[1,a)]　馬場 隆寛[2]

概要：業績は複数の要素によって変化する．その上で蓄積されたデータを分析して早期に業績の傾向をつかむことが求められている．有価証券報告書には, 企業情報をはじめとして財務・経営状況に関する様々な情報が記載されている．これらのテキストデータを活用することができれば効率的に成長企業の予測および選定が可能になると考える．本研究では, 企業業績の予測を成長・それ以外の二値分類で判断できるモデル作成に取り組む．その上で経営・財務情報が記載されたテキストから成長企業を選定できる業績予測アプリの開発を目指す．

キーワード：有価証券報告書, 業績予測, 自然言語処理

Creation of a Text Classification Model using Securities Report and development of Performance Forecasting Application

Mahisa Hoshino[1,a)]　Takahiro Baba[2]

Abstract: Performance varies depending on multiple factors. In addition, it's required to analyze the accumulated data to grasp the trend of business performance at an early stage. Securities reports contain a variety of information on financial and business conditions, including corporate information. If we can utilize these textual data, we can efficiently forecast and select growing companies. In this study, we will create a model that can predict the performance of companies using a binary classification of growing and non-growing companies. We then aim to develop a performance forecasting application that can select growing companies from texts containing management and financial information.

Keywords: Securities Report, Performance Forecasting, Natural Language Processing

1. はじめに

常に変化が求められる社会の中で, 企業は業績に対して経営戦略の策定や見直し, 改善活動に取り組んでいる．また国内最大手である信用調査会社, 帝国データバンクの調査によると 2023 年度（2023 年 4 月 1 日〜2024 年 3 月 31 日）の倒産件数は 8,881 件 [1] であり, 倒産件数の増加率も過去 30 年で最も高い 30.6 %となっている．現在でも多くの企業が倒産している状況がある．業績は複数の要素によって変化するため, 蓄積されたデータを分析して早期に業績の傾向を掴むことが求められている．

また近年では, 新しい事業への参入や既存事業の拡大を目指す合併と買収（M & A）が盛んに行われている状況である．機械学習技術の急速な進化により, M & A 候補先企業の選定を効率的に行う取り組みも進められている．今後はよりテキストおよび数値データに基づいた効果的な企業の候補先選定が加速していくと考える．有価証券報告書は公開されているデータであり, 企業情報を詳細かつ効率的

[1] 久留米工業大学 工学部 機械システム工学科
Kurume, Fukuoka 830-0052, Japan
[2] 久留米工業大学 工学部 情報ネットワーク工学科
Kurume, Fukuoka 830-0052, Japan
[a)] m211149hm@kurume-it.ac.jp

に取得することができる．これらの情報を活用して, 社会に適用する仕組みが求められている．

本研究では, 有価証券報告書を活用して企業の業績を成長する・しないの二値で判断できるテキスト分類モデルを作成する．モデルについては性能の評価を実施する．そのために各企業の経営・財務データを収集する方法について調査を行う．取得したデータから早期に業績の傾向を掴むことができれば, 成長傾向の高い企業を短時間で選定できるようになる．

2. 関連研究

有価証券報告書に関する研究やニュースなどのテキスト情報を活用した研究は多く取り組まれている．

岡島ら [2] は, 決定木を用いた M & A 候補先企業の推薦を行う研究を行った．M & A 候補先選定は人手・時間と労力のコストがかかるため, 自動で有効な候補先選定を行うことが求められており, 研究も進められている．複数の要素が作用すると考えられる M & A 候補先選定において, この研究では過去の M & A 取引事例と有価証券報告書の財務・非財務データを利用して分析を行っている．また実際に M & A が行われた上場企業の組み合わせを正例, ランダムな上場企業の組み合わせを負例として設定した．データについては実際に行われた M & A 取引を 138 件, ランダムペアによるものを 138 件, 計 276 件を使用している．正解率は 78 %の精度で正例および負例の分類が可能となっている．

土橋ら [3] は, 有価証券報告書を使用して, ESG 関連文を抽出するモデルを作成する研究に取り組んでいる．ESG とは, Environment（環境）・Social（社会）・Governance（ガバナンス）のイニシャルをとった言葉であり, 3 つの観点への取り組みや配慮が重要であることを表す．資産運用分野においては, 従来から考慮されてきた財務情報に加えて ESG 課題を考慮して投資を行う「ESG 投資」が世界的な潮流となっている．この研究では, 有価証券報告書の経営方針項目及び事業等のリスク項目の文に対してアノテーションを行うことで ESG 関連文のデータセットを作成している．各言語モデルごとに精度の比較が行われ, F 値にて評価が実施された．ファインチューニング有りの BERT を使用した場合が最も精度の良い ESG 関連文を抽出している．またグラフによる予測結果の可視化も行われている．

新津ら [4] は, ニュースのテキスト情報と株価の数値データから株価を予測する研究を行った．この研究では, 任天堂，キーエンス，日本電産の各社株価データおよびニュースを取得しデータセットを作成している．テキスト情報を分析するために BERT が用いられており文章の埋め込み表現を獲得している．その後ニュースデータと株価データを時系列なデータとして組み合わせることで LSTM に入力し学習が行われた．この研究では, 平均二乗誤差による各種モデルの比較を行った際に, BERT と LSTM を利用した平均二乗誤差の方が大きくなるという結果であった．

先行研究の調査より, テキスト情報を活用した様々な予測・分類が行われていることが読み取れる．しかし, データを活用した予測を誰もが行えるシステムの開発については新規性の余地があると考える．

これらの先行研究をふまえて, 本研究では, 有価証券報告書の経営成績・財務情報から企業の業績を成長する・しないの二値分類で判断できることを目指す．そのうえで業績予測アプリケーションの開発に取り組むことで, 実務を想定したシステム構築が可能となる．経営成績・財務情報を活用して成長企業を選定する点, 作成したモデルを活用した業績予測アプリの開発, これらは本研究の新規性であり, 強みである．

3. 提案手法

本研究では, 有価証券報告書において前事業年度の「経営者による財政状態、経営成績及びキャッシュ・フローの状況の分析」に関するテキストをもとに当事業年度の成長する・しない企業を予測する．二値分類モデルを作成することで, 当事業年度のテキストをもとに次年度の業績予測が可能となる．システム構築の流れとして, 各企業における財政・経営テキストの収集, データセットの作成, モデルの学習および評価の順に進めていく．

4. 実験方法

4.1 データセットの作成

はじめに金融庁が提供している電子開示システム EDINET[5] より各企業のデータ収集における手順について調査を行った．EDINET では EDINET API[6] を使用することにより効率的にデータを取得することが可能となる．データ取得日は 2024 月 11 月 12 日に実施し, 収集期間は 2023 年 1 月 1 日から 2023 年 12 月 31 日とした．

次に成長企業の選定を行うため, 財務体質の健全性を測る自己資本比率と, 自己資本をもとにどの程度利益をあげているかを見る自己資本利益率，２つの指標 [7] を使用する．前事業年度と比較して当事業年度の自己資本比率および自己資本利益率が, いずれも増加していれば成長する企業としてクラス 1, そうでなければ成長しない企業としてクラス 0 とする．テキストおよび数値の記載がないものについては件数から除去した上で計 3,240 件のアノテーションを行った．アノテーションとはデータに対して追加の情報タグ（メタデータ）を付加する作業 [8] のことを指す．また 3,240 件よりクラスの比率を同じ割合にするために, 書類管理番号（docID）の昇順に並び替え, 少数派クラスと同じ割合になるようにデータ削減を行った．削減後の件数は計 2,010 件となった．各クラスの件数は 1,005 件ずつとして使用する．

分類に使用する特徴量の定義を表 1, アノテーションに必要な情報を表 2 として示す.

表 1: 特徴量の定義
Table 1 Definition of features.

特徴量名	定義
経営者による財政状態、経営成績及びキャッシュ・フローの状況の分析	経営成績・財務情報が記載された文章

表 2: アノテーションに必要な情報
Table 2 Information needed for annotations.

情報	定義
自己資本比率	(純資産 - 新株予約権) ÷総資本× 100
自己資本利益率	当期純利益÷ (純資産-新株予約権) × 100
自己資本比率差	当事業年度-前事業年度
自己資本利益率差	当事業年度-前事業年度
成長企業	比率・利益率差ともに増加:1 それ以外:0

4.2 モデルの学習

作成したデータセットをもとに層化 K 分割交差検証 [9] による評価を行った. 層化 K 分割交差検証では訓練・評価データの分割を複数回行い, その平均をもって評価を行う. 特徴として, 目的変数の比率がなるべく元のままになるよう分割できるため, モデルの信頼性が向上する. 今回は, データを 5 分割として行った. また分割後のデータ件数とその割合については表 3 に示す.

表 3: 分割後のデータ
Table 3 Data after splitting.

データ	件数	分割割合
訓練データ	1,608 件 (0:804 件, 1;804 件)	80 %
評価データ	402 件 (0:201 件, 1;201 件)	20 %

本研究では, 経営成績・財務情報に関するテキストを扱うにあたり, BERT[10] を利用する. BERT とは自然言語処理のための深層学習モデルであり, 文章を文頭・文末の双方向から学習することで文脈を読むことができる特徴がある. 様々な事前学習済みモデルやデータセットが公開されているサイト Hugging Face[11] より, 日本語の長文に対応した事前学習済みモデル ku-nlp/deberta-v2-tiny-japanese-char-wwm[12] を使用する. このモデルを使用してテキストのトークナイズおよび Trainer クラス [13] を使用しファインチューニングを行う. ハイパーパラメータの設定として num train epoch=50, learning rate=2e-5, warmup ratio=0.1, weight decay =0.001, lr scheduler type =”cosine”, fp16 =True として学習を行った.

4.3 評価指標

成長する・しない企業について予測性能を評価するために, 正解率を表す Accuracy, 適合率を表す Precision, 再現率を表す Recall, 適合率と再現率の調和平均をとった値である F1 Score, この 4 つの指標指標 [14] と混同行列を使用した.

5. 実験結果

今回は 5 分割交差検証となるため, 各評価指標の平均値を算出する. また混同行列については各 Fold の合計値とする. モデルにおける各指標の結果をまとめたものを表 4 として, 混同行列の結果を表 5 に示す.

表 4: 各指標の平均結果
Table 4 Average results for each indicator.

正解率 (Accuracy)	0.659
適合率 (Precision)	0.641
再現率 (Recall)	0.740
F 値 (F1 score)	0.684

表 5: 混同行列の結果
Table 5 Confusion matrix result.

	予測: 成長しない (陰性)	予測: 成長する (陽性)
正解: 成長しない (陰性)	581	424
正解: 成長する (陽性)	261	744

表 4 では正解率が約 0.659 となっている. 適合率と再現率の関係を見ると適合率が再現率に比べて低くなっている. これは, 誤って成長するクラスに分類された件数が多いことがわかる. 適合率と再現率の調和平均をとった F 値は約 0.684 であった. また, 表 5 については正解のクラスと, それに対して予測したクラスの関係を混同行列で表したものである.

6. 考察

表 4 の実験結果より, 有価証券報告書のテキスト情報を利用することで約 0.659 の正解率が得られた. しかし, 適合率と再現率の関係では適合率が再現率に比べて低くなっている. 今回の実験を通して全体の精度も上げていくためには, 大きく 3 つのことに取り組む必要がある.

1 つ目はデータのアノテーションにおいて, 自己資本比率と自己資本利益率の両方が増加した場合と定義した. しかし, 複数の条件を満たすという観点から, 分類時の判断が難しかったのではないかと推測する. アノテーションの基準について別の指標を活用することも考える必要がある. また今回は 3,000 件以上のデータに対してアノテーションを行ったが, より効率的な方法でラベル付けを行えるように工夫していきたい.

2 つ目は, 今回特徴量として使用した経営成績・財務情報は長文のテキストである. 現状では BERT における, 一度に処理できる文章の長さが 512 トークンまでのモデルも多い. また長文のテキストを大量に学習させる場合, 学習時間が非常にかかることも判明した. 今後の精度向上に向けて, 文章の重要な部分を抽出・要約したり, 短文に区切りながら BERT に入力する工夫が必要である. 長文のテキストに対する様々なアプローチの仕方も考えていくことが求められる.

最後に 3 つ目は, EDINET API を活用することで経営・財務にかかわる文書だけではなく, 研究開発活動などの企業が取り組む将来性にかかわるテキストも収集することができることが判明した. そのため今後は, より長期的な業績予測にも活用できる可能性がある. これらのテキストも特徴量として組み合わせながら活用することで, 数年先まで信頼性の高い成長企業を選定できると考えられる.

7. まとめ

本研究を通して, 公開されている有価証券報告書から業績にかかわるテキストの収集, そして次年度の成長企業を一定程度予測できる可能性を示すことができた. 分類精度の向上は必要であるが, 作成したモデルをもとに, 業績予測が可能となった.

また EDINET API を活用することで様々なテキスト・数値情報を取得できることも分かった. 必要な情報のみを選択してデータ収集することも可能であると考える. 今回は 1 年間の企業情報を収集したが, 一定の企業に絞り, より長期間のデータを取得することで信頼性の高い企業特化型の業績予測も可能になると考える.

一方で, モデルの分類精度をより向上させるためにはデータの選定基準や, 長文テキストの学習方法, テキストデータ同士の組み合わせなど, 課題も多く見つかった. 今後は, グラフや図など可視化できる機能を備えてユーザーがより直感的に理解しやすい仕組みを目指す必要がある.

今後は, 実用性のあるシステム構築にも取り組んでいくとともに, データの効率的な収集方法や業績の分類精度向上につながる方法を探していきたい.

参考文献

[1] 株式会社帝国データバンク：倒産集計 2023 年度｜株式会社 帝国データバンク [TDB], 株式会社 帝国データバンク（オンライン）, 入手先 ⟨https://www.tdb.co.jp/report/bankruptcy/aggregation/4z0yynojmp/⟩（参照 2025-02-11）.

[2] 岡島右馬, 許子微, 市瀬龍太郎：決定木を用いた M&A 候補先企業の推薦, 人工知能学会第二種研究会資料, Vol. 2023, No. FIN-031, pp. 150–153 (2023).

[3] 土橋諒太, 中田和秀：BERT を用いた有価証券報告書からの ESG 関連文抽出, 人工知能学会第二種研究会資料, Vol. 2021, No. FIN-026, p. 09 (2021).

[4] 新津康平, 吉浦紀晃：BERT と LSTM を利用した株価予測, 研究報告数理モデル化と問題解決 (MPS), Vol. 2021, No. 5, pp. 1–6 (2021).

[5] 金融庁：EDINET, 金融庁 (online), available from ⟨https://disclosure2.edinet-fsa.go.jp/week0010.aspx⟩ (accessed 2025-02-11).

[6] 金融庁：EDINET サインアップまたはサインイン, 金融庁（オンライン）, 入手先 ⟨https://api.edinet-fsa.go.jp/api/auth/index.aspx?mode=1⟩（参照 2025-02-11）.

[7] 財務総合政策研究所：法人企業統計からみえる企業の財務指標 : 財務総合政策研究所, 財務省（オンライン）, 入手先 ⟨https://www.mof.go.jp/pri/reference/ssc/zaimu/index.htm⟩（参照 2025-02-11）.

[8] NTT Communications Corporation: アノテーションとは？意味・定義 — IT 用語集 — docomo business Watch — ドコモビジネス — NTT コミュニケーションズ 法人のお客さま, NTT Communications Corporation（オンライン）, 入手先 ⟨https://www.ntt.com/bizon/glossary/j-a/annotation.html⟩（参照 2025-02-11）.

[9] codexa チーム：交差検証（Python 実装）を徹底解説！図解・サンプル実装コードあり, codexa.net（オンライン）, 入手先 ⟨https://www.codexa.net/cross_validation/⟩（参照 2025-02-11）.

[10] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019).

[11] Hugging Face: Hugging Face – The AI community building the future., Hugging Face (online), available from ⟨https://huggingface.co/⟩ (accessed 2025-02-11).

[12] Language Media Processing Lab at Kyoto University: ku-nlp/deberta-v2-tiny-japanese-char-wwm, Hugging Face (online), available from ⟨https://huggingface.co/ku-nlp/deberta-v2-tiny-japanese-char-wwm⟩ (accessed 2025-02-11).

[13] 我妻幸長：BERT 実践入門 PyTorch+Google Colaboratory で学ぶ新しい自然言語処理技術, 株式会社 翔泳社 (2023).

[14] OPTiM TECH BLOG:【初心者向け】機械学習におけるクラス分類の評価指標の解説 - OPTiM TECH BLOG, OPTiM Corp.（オンライン）, 入手先 ⟨https://tech-blog.optim.co.jp/entry/2021/05/31/100000⟩（参照 2025-02-11）.