有価証券報告書へのテキストマイニングの適用に関する文献レビュー

近藤隆史
石光　裕

要　旨

本稿では，有価証券報告書の分析にテキストマイニングを適用した先行研究を広く収集し，研究の目的や分析項目・単位，手法などの観点から整理することによって，テキストマイニング適用の意義と課題について検討している．

1　はじめに

本稿では，有価証券報告書（以下，有報）のテキスト形式の定性情報[1]にテキストマイニング（Text Mining; TM）を適用した先行研究をレビューし，会計分野において，有報にTMを適用する意義や課題について検討する．

近年，ニュース記事やソーシャルメディアへの投稿記事だけでなく，企業のIR情報などから，金融関連の膨大な定性情報がテキスト形式で容易に入手できるようになり，それら金融テキストデータへのTMが学術目的でも多く用いられてきている．その中でも有報は，会計分野に限らず[2]，経営や工学分野でも，TMの適用対象として広く関心を集めつつある[3]．ただし，TMが適用される目的や方法は研究により様々であり，その動向などを整理しておくことは今後の研究にむけて有用であると思われる．

そのため，本稿では，有報およびTMの特徴を概観した上で（2節），先行研究をもとに，研究者の関心は有報のどこにあるのか，TMがどのように使われ，どういった知見が得られているのかについて整理する（3節）．対象とした文献は，学術分野を問わず広く収集している．それらレビューにもとづき，会計分野における有報のテキストデータにTMを適用する意義や課題について検討（4節）を行う．最後に，考察を取りまとめ本稿をしめくくる．

1）　投資家による適切な投資判断のための財務情報以外の情報は記述情報と呼ばれており金融庁（2019），有報で開示される経営方針やリスク情報といった定性的な内容もこれに含まれる．本稿では，記述情報をその特性に着目して定性情報とし，定性情報のデータ形式を指す場合は，テキストデータとしている．

2）　例えば，有報の分析にTMを適用した会計研究の状況ついては，雑誌『企業会計』（2022年2月号）でも特集されているので参照されたい．

3）「有価証券報告書」と「テキストマイニング」を検索ワードにGoogle Scholar（https://scholar.google.co.jp）でウェブ検索したところ，2006年ごろから論文が散見されるようになり，直近5年では，2018年で13本，2019年で14本，2020年で20本，2021年で25本，2022年24本のように徐々に増える傾向にある．

2 有価証券報告書とテキストマイニングの特徴

ここでは，本稿での議論に必要な範囲で，有報および TM について触れておこう．まず，有報は，金融商品取引法によって金融商品取引所（証券取引所）に株式公開している会社などに対して提出が義務付けられている，企業の状況を外部に報告する書類である．その内容は，大きく「第一部 企業情報」,「第二部 提出会社の保証会社等の情報」，そして，「監査報告書」の 3 つからなっているが，第一部の企業情報が分量のほとんどを占めている．

それでは有報には，TM の対象となり得るテキストデータがどれくらい含まれているのだろうか．図表 1 は，2021 年 3 月決算企業 2,366 社の有報を対象に，「第一部 企業情報」の代表的な項目ごとの平均文字数を示したものである [4]．その多い順に,「経理の状況」,「提出会社の状況」,「事業の状況」となっており，これらは 1 万字を超えている．「経理の状況」の文字数が突出して多いのは，注記のテキストデータが多いためと考えられる．

また，詳細な項目分類（以下，詳細項目）からは，「コーポレート・ガバナンスの状況等（以下，ガバナンス状況）」の文字数が最も多く，次いで「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（以下，MD&A）」で，それ以下,「株式等の状況」,「事業等のリスク（以下，事業リスク）」,「経営方針，経営環境及び対処すべき課題等（以下，対処すべき課題）」などが続いており，いずれも多くの定性情報が含まれている．これらの項目では，その定性情報の多くがテキスト（単語や文章）で構成され，内容に応じて細かく分けて記載されている．

4） 文字数は，XBRL タグ（次頁参照）の要素名 ix:nonNumeric を指定してカウントした．また「企業の概況」の平均文字数とその詳細項目（「主要な経営指標等の推移」,「沿革」など）の平均文字数の合計が異なるのは，いくつか省略したマイナーな詳細項目があるからであり，他項目でも同様の表示となっている．なお,「経理の状況」については，企業ごとに設定された項目が多いため，詳細項目を省略した．

図表 1　有価証券報告書（2021 年 3 月）の「第一部　企業情報」の項目ごとの記載文字数

項目	平均文字数	詳細項目	平均文字数
企業の概況	7,363	主要な経営指標等の推移	2,394
		沿革	1,700
		事業の内容	1,379
		関係会社の状況	1,305
		従業員の状況	571
事業の状況	15,788	対処すべき課題	2,910
		事業リスク	4,232
		MD&A	7,350
		経営上の重要な契約等	331
		研究開発活動	921
設備の状況	1,951	設備投資等の概要	257
		主要な設備の状況	1,362
		設備の新設，除却等の計画	281
提出会社の状況	32,232	株式等の状況	4,924
		自己株式の取得等の状況	1,188
		配当政策	520
		ガバナンス状況	25,468
経理の状況	41,086	—	—
提出会社の株式事務の概要	583	—	—
提出会社の参考情報	690	—	—

また有報は，金融庁が運用する EDINET（Electronic Disclosure for Investors' NETwork）と呼ばれる電子情報開示システムにおいて無料で閲覧でき，データを入手することができる．データには，XBRL（eXtensible Business Reporting Language）形式のものが含まれる．XBRL では，勘定科目や項目名などの要素名や金額，日付などの属性がタクソノミ[5]として定義されており，これにしたがって，情報を識別するためのタグ（XBRL タグと呼ぶ）がつけられている．そのため XBRL タグにより構造化[6]されたテキストは，コンピュータで解析がしやすく，簡単なプログラミングで必要なデータを抽出することができる．また，EDINET からは，Web API（Application Programming Interface）が提供され，プログラム上で有報データを効率的に取得できる．このように有報のテキストデータは，コンピュータ上で扱いやすい形式になっており，インターネットを介して入手しやすくなっている．

5)　タクソノミとはコンピュータ処理を前提とした電子的なタグ集合のことで，金融庁が定義した EDINET タクソノミと提出者により作成されるタクソノミに分けられる（本稿では両者を区別せず用いている）．XBRL 形式の有報も，それらタクソノミに従い，インスタンスと呼ばれるタグ付けされた実データのファイルから作成されている．

6)　一般的に，構造化されたデータといえば，専用のツールで構造をパースでき，必要なデータ（値）を抽出できるような形式を指すが，有報も記載の項目のレベルで構造化されていると言える．ただし，項目内では，発信者による自由形式で記載されていることも多く，その点では，半構造的なデータ（要素）も併せ持っているといえる．

以下，本節の残りで TM について説明しよう．TM は，テキストから意味のある情報をマイニング（探索）することを意味する．具体的には，コンピュータ処理[7]を前提とし，自然言語処理の技術を用いて，テキストデータを単語などの小さな単位に分割した上で，データ解析の手法により定量的に判断を行う一連の手続きを指している（和泉ほか 2022; 金 2021）．

自然言語処理の技術には，テキストを形態素と呼ばれる意味をなす最小の単語単位に分割する（分かち書きと呼ばれる）形態素解析，分かち書きされた形態素間の文法上の係り受け関係を明らかにする構造解析，単語や文の言語上の意味を同定する意味解析，さらに，因果を含む文や文と文の間の関係を明らかにする文脈解析が含まれる（和泉ほか 2022）．ただし，こうした自然言語処理を行うには，まずテキストを何らかの小さな単位に分かち書きする必要があり，TM では，分かち書き，つまり形態素解析を分析の起点とすることが多い[8]．そのため，意味解析でも文脈解析でも，それら言語上の意味や文脈そのものを分析対象にはできず，分かち書きされた単語に依拠しているのが現状である（金 2021）．

有報のテキストデータも自然言語処理（形態素解析）が施され，コンピュータが扱いやすいデータ形式に変換される．その変換されたデータに，一般的なデータ解析と同様の手法が適用される．データ解析といっても，TM では，単語の語彙（種類）やその頻度，単語間の関係（共起）といったシンプルな方法でテキストの特徴が捉えられることが多い．より高度な手法として，回帰分析や主成分分析，対応分析などの多変量解析のほか，トピックモデルやサポートベクトルマシン（SVM）といった機械学習，さらに，Word2Vec[9] などのニューラルネットワークの技術を応用した学習モデルまで多岐に渡っている．

こうした TM の一連の手続きは，有報を含む金融テキストの分析には欠かせない方法の一つとなっている（和泉ほか 2022）．次節では，先行研究のレビューを通して，TM が有報のテキストデータにどのように適用されているのか概観する．

3 有価証券報告書を対象とした先行研究

本稿では，有報に TM を適用した文献を広く収集している．その収集の手順として，CiNii Research[10] にて，検索ワードを “ テキストマイニング ”（または “ テキスト分析 ”）と “ 有価証券報告

7） TM には，Python（https://www.python.org）や R（https://www.r-project.org）などのプログラミング言語や KH Coder（https://khcoder.net）などの専用アプリケーションが用いられることが多い．

8） 自然言語処理において，単語に分かち書きした上で意味を理解する場合，同じ文脈に出現する単語は近い意味をもつといった仮定をおく．実装上は，単語の意味は，その単語の周辺に出現しやすい別の単語と関連付けて理解されることが多い．

9） 2013 年に Google の研究者トマス・ミコロフ氏によって提案された，単語の意味をベクトル表現した分散表現を生成するための学習モデルである．単語の類似度などがベクトル計算で出せるのが特徴である．

10） https://cir.nii.ac.jp

書 " として検索し，内容を確認した上で，関連ある文献をさらにたどる作業を繰り返した[11]．TM の手続きは，統計的方法のように標準化されているとはいえないため，本稿では，有報のテキストデータ中の単語に注目し，それらを変換した特徴量を分析に用いている研究レビューを対象とした[12]．また，TM の技術は進展も早く多岐に渡っており，こうした動向を捉えるため，学会誌の論文だけでなく，大学紀要論文や学会報告要旨集に収められた文献なども広く含めた．結果，本稿で取り上げた先行研究は，2006 年から 2022 年までの 56 本だが，そのうちの 41 本が直近 5 年に集中している．一覧は付録としている．

3.1　サンプル

付録の「サンプル」の列は，分析に利用された有報の決算期やサンプル数を示している．先行研究では，その時点で取得できる有報全てが分析対象となっているわけではない．特定の年度や決算期，さらに，TOPIX や日経銘柄など特定の企業グループ，さらに，単一の企業に絞ったものが散見される．例えば，特定のイベント前後（新型コロナウィルスや組織改革）の比較では，その前後の年の有報に限定していたり，ある特徴文を自動抽出するための学習モデルの予測精度を向上させるために少量サンプル[13]を用いている研究もある．

3.2　分析対象となっているテキスト

研究者は，TM の対象として，有報のテキストデータのどの部分に関心をもっているのだろうか．先の図表 1 では，項目ごとの平均文字数が示されていたが，記載される文字数の多い箇所へ研究者の関心が集まることは容易に予想される．図表 2 は詳細項目ごとの先行研究の数（1 文献につき複数項目のカウントあり）を示している．

先行研究の件数は，平均文字数の多い詳細項目を中心に行われているが，最も文字数の多い経理の状況のテキストデータを対象とした研究は少なかった．これは，そのテキスト部分のほとんどが注記であり，多くは決められた単語と数値の組み合わせによるもののため，定量情報と同様に扱うことができ，TM の対象とはなっていないためと考えられる．一方，件数が最も多いのは事業リスクであった．企業の事業リスクに関する情報は，投資家にとって重要なのは言うまでもなく，将来業績にも影響すると考えられており，分野問わず研究者の関心は高いのかもしれない．

11）　文献タイトルには先の検索キーワードが含まれていないが TM を適用している先行研究もあり，これらをできるだけ多く取り入れるためにこのような手順としている．

12）　有報のテキストデータを使用した研究であっても，注記の項目数に着目した中島・音川（2020），注記に含まれる項目別の金額に着目した澁谷（2018）など，単語や文を分析の単位とはしていないものはレビューから除いている．

13）　こうした試みには，高野ほか（2022）などがあげられる．

図表 2 詳細項目ごとの先行研究の件数

分析対象となった詳細項目（項目）	平均文字数（2021.3）（図表 1 より）	先行研究の件数
事業リスク	4,232	22
対処すべき課題	2,910	11
MD&A	7,350	10
ガバナンス状況	25,468	6
配当政策	520	3
経理の状況	41,086	2

3.3 分析の単位と手法

有報のテキストデータがどの単位で分析されているのかについて, 付録では,「単語レベル」と「センテンスレベル」に分類している．まず，単語レベルとは，TM による分析結果が，分かち書きされた単語の単位に基づいている研究を指している．一方，センテンスレベルは，文（文章）を分析の単位とした研究を指している．実際の分析には分かち書きが必要だとしても，目的が文の意味の分析や特徴文の抽出にある研究である．なお, 解析器[14]やそのための辞書[15]などの実装状況ついても文献から分かる範囲で示している．

また，TM で使われる特徴量や手法も示している．特徴量については，異なり語の割合を測った語彙指標の TTR（Type Token Ratio），単語の文書内での重要度の指標の TF-IDF（Term Frequency - Inverse Document Frequency）がよく使われる．最近では，記載内容の質的な面を評価するための可読性[16] や単語のトーンあるいはセンチメント（ポジティブ，ネガティブ，あるいは中立といった感情を示す極性）などにも関心が寄せられている．さらに付録には，特徴量に用いられた単語の品詞も示している．手法については，データ解析の主だった方法を示している．

3.4 研究のタイプ

文献リストの「タイプ」では，研究のタイプが記述，予測，学習の 3 つに分類されている．記述とは，記載内容の特徴や傾向を明らかにする研究を指している．このタイプの先行研究では，有報の開示状況の把握だけでなく，分析者の依拠する理論的な枠組みや概念の下で企業行動についての

14） プログラミング環境でよく使われるものとして，形態素解析器には MeCab（https://taku910.github.io/mecab/）や JUMAN （https://nlp.ist.i.kyoto-u.ac.jp/?JUMAN），ChaSen（http://chasen-legacy.osdn.jp）など，構造（係り受け）解析器には CaboCha（https://taku910.github.io/cabocha/）などがある．

15） 日本語テキストの形態素への分かち書きには，解析器のほか単語を登録した辞書が必要になる．解析器 MeCab や ChaSen などで推奨されるシステム辞書として，IPA コーパス（情報処理振興事業協会）に基づく IPA 辞書（ipadic）ほか，国立国語研究所による UnDic（https://clrd.ninjal.ac.jp/unidic/）やシステム辞書を拡張するためのユーザー辞書として使われる NEologd（https://github.com/neologd）などがある．また，必要に応じてより専門的なユーザー辞書がオリジナルで作成されるなどしている．

16） 文書の可読性は総文字数および文の難易度（Fog Index）で測定されている（Li 2008）．

考察も試みられている．

つぎに予測とは，有報の記載内容とそれと関連がありそうな量的データをリンクさせた分析を指す（金 2021）．実際には，有報の各項目の特徴量を説明変数とし，将来業績などの財務指標を目的変数とした回帰モデルを検証する研究が散見される[17]．将来業績を予測する上で，どういった特徴量が有意に影響するのかは，現状で共通の認識は得られていないものの，最近ではテキストの可読性の指標が注目され，その有効性が検証されている．また，予測タイプの中でも，学術的な仮説検証を明示的に行っている文献にはその旨をリストに記している．

最後に，学習とは，特定のタスク遂行のための学習モデルの構築（提案）を試みる研究を指している．モデルの構築には，学習データと評価データを用意した教師ありの学習が取られることが多い．先行研究では，文脈を判断するための手がかり語を何にするかなどの工夫をすることで，有報のテキストから，経営状況に影響を及ぼした要因や因果関係の内容を含む特徴文を，如何に精度高く抽出するかが検討されている．こうしたタスク遂行では，深層学習が主流であり，先行研究では，BERT[18] による言語学習モデルがよく利用されていた．

4　有価証券報告書の分析に TM を用いる意義と将来の展望

前節でのレビューをふまえ，本節ではまず有報の分析へ TM を適用することの意義について検討したのち，会計研究における今後の課題を提示する．

TM 適用の第 1 の意義は，人間の主観や直感に依存せず，定性情報を扱うことができる点にある．コンピュータ処理を前提とした TM を適用することで，まずテキスト（文書）が単語分割（分かち書き）されたデータに変換され定量データと同じように扱えるため，既存のデータ解析の手法によって定量的な分析が可能になる．これは有報に TM を適用した場合にも当てはまる．コンピュータの高いデータ処理能力により，膨大な有報テキストデータを効率よく処理することで，分析の客観性を保ちながら，分かち書きされた単語（形態素）の中から，興味深い関係・パターンの発見が期待できる．

第 2 に，TM により，有報テキストの構造的な特性が持つ利点を引き出すことができる点がある．先述の通り，有報のテキストデータはタクソノミ（XBRL タグ）により構造化されている．こうしたデータにより，研究者は有報のテキストデータ全体の中から，自らの関心と関係する部分のみを

17）反対に有報の各項目の特徴量に影響する要因（企業属性）を検証する研究も予測タイプに含めている．

18）Bidirectional Encoder Representations from Transformers の略で，2018 年に Google が開発した汎用言語学習モデルである．従来のモデルとは異なり，深層学習モデルの一つ Transformer の仕組みを使って文章を文頭と末尾の双方向から事前学習することで，単語の文脈が理解できる（例えば，同じ単語でも文脈に応じた意味の違いが識別可能）といった性能の高さと，汎用的な学習済みモデルに対しタスクに応じたチューニングを施すことにより学習の転移性もあることから，現在では様々なタスクに利用されている．

抽出して分析できる．研究者は，TM の分析結果を，有報の記載内容の意図と対応させることで，何らかの意味づけ・解釈が可能になる[19]．TM を使うことで，有報のテキストの特性を活かしながら，大量のデータを効率的に処理できる．先行研究では，有報のテキストデータは，開示情報の内容把握だけでなく，研究者が依拠する概念や理論を背景に意味付けをしたうえで，企業の経営行動の分析にも使われている[20]．

これらの意義をふまえて，今後取り組むべき課題を示す．第 1 に，有報の項目間の関係性を捉えた分析の必要性についてである．先行研究の多くは，有報内の複数の記載項目を分析の対象とするものの，その分析は項目ごとに独立に行われ，複数の項目間の関係についてはほとんど扱われていない．例えば，事業リスクを分析対象とすると，先行研究でもあるように，発信者（経営者）が認識するリスクの種類やトピックを明らかにできる．一方で，認識されたリスク要因が対処すべき課題やガバナンス状況といった他の項目でどう扱われているのか（例えば，どの程度同様の単語が出現するのか，あるいは他の項目から事業リスクの項目がどの程度参照されているのか）は，企業のリスク対応を総合的に評価する際には重要となるが，こういった検証は行われていない[21]．今後，有報の分析に TM を適用する上で，有報内の項目間の関係性をどう定量的に捉えるかに加え，そうした関係性が企業にどのように影響するのかの探索が必要になるだろう．

分析対象の拡大が必要なのは有報内の項目間に限らない．有報には，業績変化についての説明，事業リスクへの関心の高さや課題への取り組み姿勢などを示唆する文章が多く含まれるが，これらの定性的な情報が実質的にどのような影響をもつかは，客観的な（有報の）外の指標と結びつけて，評価する必要がある．探索タイプの先行研究では，仮説検証を含め，有報の定性情報がその外の指標に及ぼす影響についての検証が試みられているが[22]，さらなる蓄積が求められる．

第 2 に，分析の単位を検討する必要性が指摘できる．その候補として，一つに，単語ベースであったとしても，その語彙やその頻度だけでなく，言語上の性質に注目するものと，もう一つ，分析単位を単語から文節や文章にまで拡大するものとが考えられる．

前者の例としては，単語の極性（肯定的か否定的か）を評価するトーン分析が挙げられる．トーン分析は先行研究でも少ないが，有報の中での環境やリスク，業績といった経営（事業）の状態に

19） 言うまでもなく，TM からの結果そのものが何か特定の意味を発するものではない．有報のように，タクソノミ（XBRL タグに付与された要素名）により，TM の分析からの結果を補完できる．例えば，事業リスクの記載の中で，「原料価格」が出現すれば，発信者（経営者）は，その高騰などを自社のリスクとして認識していると考えるのが妥当な解釈だろう（もちろん，「原料価格」と合わせて，「変動」や「高騰」などの単語との共起も見るのが望ましい）．

20） 近藤・石光（2016, 2020）では，有報のガバナンス状況から内部統制に対する経営者の意識が測定され，中山・津田（2018）では，事業リスクから企業の ISO マネジメントの実施（行動）を捉えようとしている．他にも，喜田（2006）では，有報のテキストデータをもとに組織の認知が扱われている．

21） 目的は異なるが，有報の中で他の項目への参照を扱った研究に首藤・緒方（2008）があげられる．

22） 廣瀬ほか（2017）では，有報のテキストの可読性が将来業績に及ぼす影響が検証され，また，佐藤ほか（2021）や Kim and Yasuda（2018）では，事業リスク関連の単語と外部の指標（株式収益率に伴うリスクや企業の総リスク）との関連が検証されている．

対する経営者の評価の分析が試みられている[23]．こうしたトーン分析では多くの場合，単語の極性は，専用の辞書（極性辞書）[24]を使い分類され，テキストの極性の度合いが評価される．一般に，極性辞書の作成には，大量の単語への極性の付与が必要になるが，どういった文書を参照したかによって単語と極性の対応は異なる．先行研究もそれほど多くなく，トーン分析の手法を洗練させていく上で，分析に不可欠な極性辞書にも改善の余地は残されている．さらに，最近では，有報を含む金融テキストデータをベースに，極性判定のための言語学習モデルの開発が試みられている[25]．

後者で挙げているのは，分析単位の単語からセンテンスへの拡張である．先行研究では，単語レベルのものが多かったが，文章中の全ての単語が使われているわけではなく，分析に使用できる情報が制限された状態となる．センテンスレベルの研究は，学習タイプにおいてみられたが，多くで様々な特徴文[26]の識別が試みられ，そうした文章について，企業の将来業績・行動を予想する上での関連性が示唆されている．センテンスレベルの分析は，会計研究ではほとんど見られない．一方で，企業の将来業績の予想を目的とした研究は多いため，センテンスレベルの研究でのアプローチを援用する余地は十分にあるだろう．また，センテンスの分析では，言語学習モデルの開発も行われていて，その成果は，特徴文の抽出だけでなく，トーン分析にも応用できる．こうした他分野の知見をうまく援用することも必要と考えられる．

第3は，より複雑なテキストへの対応についてである．有報のテキストデータは，XBRL形式のタグ付けされているテキストでも，文字だけで構成されるプレーンテキストではないことが多い．有報のテキストデータそのものはHTML[27]形式のファイルであって，文字や文のレイアウトには，HTMLタグによる単純な段落構成のほか，表形式も多用される．こうしたタグは自然言語処理では不要であり，事前に何らかの処理を施す必要がある．ただし単純にタグを取り除くと，テーブル構造を崩してしまい，せっかく表中のテキスト（単語や語句，文章）が持っている情報が分析には取り込めなくなる．また，有報では，記載すべき内容は決まっているものの，項目（事業リスクやMD&Aなど）によっては，基本的に自由形式で書かれており，テキスト中に見出しが任意で設けられるなどしている．見出しの語句は，記載内容をより正確に捉えるための手がかりとして役立つ反面，コンピュータ処理では，見出し語句と本文のテキストの区別はつきにくく両者の判別は容易でな

23）佐久間・田中（2018）では業績の因果関係文の中の単語について，加藤・五島（2021）ではMD&Aの中の単語についての極性の定量評価が試みられている．

24）Loughran and Macdonald（2011）では，米国上場企業が提出する10-Ksのテキストに出現する単語の極性辞書が作成されている．加藤・五島（2021）や金（2022）では，Loughran and Macdonald（2011）で用いられた辞書を日本語に翻訳して有報の単語のトーン分析に利用している．最近では，日本語の金融テキストに特化した極性辞書が公開されている（https://sites.google.com/socsim.org/izumi-lab/tools）．

25）鈴木ほか（2022）などでは金融BERTモデルが提案されている．

26）例えば，対処すべき課題において未来志向の文の判別（小寺ほか2019），事業リスクにおいて，リスクを予想・仮定したものか現に直面しているものかの判別（藤井ほか2020）も試みられている．

27）Hyper Text Markup Languageの略で，ウェブページ作成用のコンピュータ言語である．EDINETからはインラインXBRLと呼ばれるHTMLファイル（拡張子はhtm）が提供されている．

い[28]．こうした表の構造や見出しの情報を正確に取り込むことができれば良いが，現在の状況では過度に手間がかかり，TM の利便性が損なわれてしまう．その点は，さらなるタクソノミの発展に期待するところでもあるが，それとは別に，TM のテンプレート化された自動処理の手続きの開発や，各要素にも何らかの形で対応できるよう検討[29]することも必要だろう．

5 おわりに

本稿では，有報のテキストデータの分析に TM を適用した先行研究を，会計分野に限定せず広くレビューし，TM を適用する意義や課題について検討してきた．レビューから，TM は自然言語処理とコンピュータの高い処理能力により，研究者のそれぞれの目的に応じて，有報の定性情報を構成するテキストデータから興味深い結果が得られるよう支援してくれる有効な分析方法であることが示唆される．

さらに，会計分野において，有報の分析に TM を適用するにあたって，いくつかの課題も提示した．具体的には，有報内での項目横断的な分析方法を探り，有報外の客観的指標との関係についての研究もさらに蓄積していくこと，トーン分析の手法を極性辞書含め洗練させること，センテンスを分析で扱えるように他分野での知見を活かすこと，さらに，複雑なテキストデータに対応するための方法を検討すること，などを挙げた．

本稿では，有報への TM の適用に焦点を絞っているため，国内の先行研究のレビューが中心となっている．いうまでもなく，海外の論文では TM による金融テキストの分析が盛んに行われている．テキストの言語の違いは，TM の実施に大きな影響を与える．そのため，海外の論文で取られている方法を有報の分析にそのまま用いることはできないかもしれないが，TM のより発展的な利用を考える上でそれらの検討は有用となるだろう．この点については，また別稿にゆずりたい．

謝辞　本研究は，2021 年度公益財団法人牧誠財団研究助成金，科研費（20K02024）による成果の一部である．

28） テキストに現れる見出し語句をどう扱うかは研究による．中山・津田（2018）では，事業リスクにおいて，リスク数を正確にカウントするため，見出しがある場合はそれを優先して一件としている．テキストの中で見出しなのか本文なのかといった一見簡単な区別も，何らかの統一的なタグ付がなければ正確な判断には目視しかない．こうした点は，TM が前提とするコンピュータ処理の難しさといえ，これに対処するテキストデータの前処理については，金ほか（2022）などを参照されたい．

29） 例えば，情報処理の分野では，有報のテーブル形式の解析に特化した取り組みもある（木村ほか 2022）．

引用・参考文献

Kim, H. and Y. Yasuda, 2018. Business risk disclosure and firm risk: Evidence from Japan, *Research in International Business and Finance*. 45: 413-426.

Li, F. 2008. Annual report readability, current earnings, and earnings persistence, *Journal of Accounting and Economics*. 45 (2): 221-247.

Loughran, T. and B. Macdonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, *Journal of Finance*, 66 (1): 35-65.

石光裕. 2018.『研究開発費情報と投資家行動』中央経済社.

和泉潔・坂地泰紀・松島裕康. 2022.『Python による金融テキストマイニング』朝倉書店.

伊藤健顕. 2019.「トピックモデルを用いた MD&A 情報の分析」日本会計研究学会第 78 回全国大会, 1-12.

伊藤健顕・越智学. 2016.「継続企業の前提に関する注記の情報量と株主資本コスト」日本会計研究学会第 75 回全国大会, 1-14.

延東晃. 2020.「サプライチェーンにおける調達リスクマネジメントに関する研究」『危険と管理』51: 21-42.

延東晃. 2022.「有価証券報告書の制度改定が企業の調達リスク開示に与える影響に関する研究」『危険と管理』53: 92-111.

大森寛文. 2014.「電機業界における経営課題の認識構造と実行動に関する知識の発見」菰田文雄・那須川哲哉（編著）『ビッグデータを活かす技術戦略としてのテキストマイニング』（第 2 章）中央経済社.

加藤大輔・五島圭一. 2021.「有価証券報告書のテキスト分析：経営者による将来見通しの開示と将来業績」『金融研究』40（3）: 45–75.

北島良三・上村龍太郎・酒井浩之・中川慧. 2019.「2016 年度決算を対象とした社是と企業業績の関係（第一報）」『人工知能学会全国大会論文集』1-4.

喜田昌樹. 2006.「アサヒの組織革新の認知的研究：有価証券報告書のテキストマイニング」『組織科学』39（4）: 79-92.

城崎渉吾・松井嗣夢・金澤威朋・伊藤有由・坂原沙月. 2021.「有報のテキストマイニングによる柔軟な勤務形態導入企業の分析」『人工知能学会全国大会論文集』1-4.

木村泰知・近藤隆史・門脇一真・加藤誠. 2022.「有価証券報告書の表を対象とした情報抽出タスクの提案」第 29 回金融情報学研究会（人工知能学会第二種研究会資料）, FIN-029, 32-38.

金鉉玉・矢澤憲一・伊藤健顕. 2022.「経営者交代が記述情報の変化に与える影響：有価証券報告書における記述情報を用いて」『会計プログレス』23: 49-67.

金明哲. 2021.『テキストアナリティクスの基礎と実践』岩波書店.

許麗梦・金明哲. 2021.「財務に関する数値データ及びテクストデータを用いた企業倒産の判別分析」『データ分析の理論と応用』10（1）: 45–57.

小寺俊哉・佐藤史仁・田中良典. 2018.「テキストマイニングによる有価証券報告書の因果関係文以外の特徴文の抽出」『日興リサーチレビュー』3: 1–15.

小寺俊哉・田中良典・佐藤史仁・佐久間洋明・坂地泰紀・和泉潔. 2019.「有価証券報告書からの未来志向文の抽出」『人工知能学会全国大会論文集』1–4.

近藤隆史・石光裕. 2016.「経営者のマネジメント・コントロールへの意識と企業業績：有価証券報告書のテキスト分析を通して」京都産業大学経営学部 Discussion Paper Series.

近藤隆史・石光裕. 2020.「マネジメントコントロールが将来業績に与える影響：コーポレートガバナンス情報へのテキスト分析の適用」『メルコ管理会計研究』12（1): 17–29.

佐久間洋明・田中良典. 2018.「有価証券報告書に含まれるテキスト情報と企業業績の関係」『日興リサーチレビュー（日興リサーチセンター）』3:1-13.

佐藤史仁・佐久間洋明・小寺俊哉・田中良典・坂地泰紀・和泉潔. 2018.「有価証券報告書からの因果関係文の抽出」『人工知能学会全国大会論文集』1–4.

佐藤隆清・池田直史・井上光太郎. 2021.「有価証券報告書のテキストマイニングによる株式のリスクファクター分析」『証券アナリストジャーナル』59（1): 99–111.

佐藤慧・酒井浩之・高野海斗・井上大輔・藤野加奈. 2021.「上場企業における企業理念と業績要因の関連性の推定」『人工知能学会全国大会論文集』1–4.

澁谷英樹. 2018.「海外との税率差がわが国の法人実効税率に与える影響：税効果会計に関する注記を用いた推計」『税に関する論文入選論文集』14: 49-89.

首藤昭信・緒方英明. 2009.「有価証券報告書における「財政状態及び経営成績の分析（MD&A）」について」『プロネクサス総合研究所 研究所レポート』3.

白田佳子・坂上学. 2008.「人工知能アプローチによる「継続企業の前提」の解析」高田 敏文（編著）『事業継続能力監査と倒産予測モデル：テキストマイニングによる非会計情報の分析』（第 6 章）同文舘出版.

白田佳子・竹内広宜・荻野紫穂・渡辺日出雄. 2009.「テキストマイニング技術を用いた企業評価分析：倒産企業の実証分析」『年報 経営分析研究』25: 40-47.

鈴木雅弘・坂地泰紀・平野正徳・和泉潔. 2021.「金融文書を用いた事前学習言語モデルの構築と検証」人工知能学会第二種研究会資料 2021（FIN-027）: 5-10.

鈴木雅弘・坂地泰紀・平野正徳・和泉潔. 2022.「事前学習と追加事前学習による金融言語モデルの構築と検証」人工知能学会第二種研究会資料 2022（FIN-028）: 132-137.

陶逸辰.2022.「重大なリスクに対する異業種間における企業の情報開示：コロナ禍の下で企業の開示実態に対する実証研究」『商学研究論集』56（2）: 239-258.

高野海斗・岡田知樹・清水裕介・中川慧. 2022.「有価証券報告書からの将来の配当政策文のテキストマイニング」『人工知能学会全国大会論文集』1-4.

高野海斗・酒井浩之・北島良三. 2019.「有価証券報告書からの事業セグメント付与された 業績要因文・業績結果文の抽出」『人工知能学会論文誌』34（5）: 1-22.

竹内広宜・荻野紫穂・渡辺日出雄・白田佳子. 2008.「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』124-127.

辻野幹実・永井義満・石津昌平. 2010.「有価証券報告書を対象とした企業の経営課題と対策の抽出方法」『経営情報学会全国研究発表大会要旨集』630-63.

土屋和之. 2018.「事業等のリスクの分析：記載内容の類似度にもとづくアプローチ」『千葉商大論叢』55（2）: 113-133.

土屋和之. 2020.「事業等のリスクの分析：記載内容のトピックにもとづくアプローチ」『千葉商大論叢』57（3）: 185-197.

土橋諒太・中田和秀. 2021.「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 2021（FIN-026）』1-7.

中島隆広. 2020.「有価証券報告書における定性的情報の記述情報量と可読性の決定要因に関する実証研究」神戸大学大学院経営学研究科大学院生ワーキング・ペーパー.

中島隆広・音川和久. 2020「税効果会計の注記における開示項目数の決定要因」『国民経済雑誌』223（3）: 31-53.

中野貴之. 2010.「財務諸表外情報の開示実態：事業等のリスクおよび MD&A の分析」山崎秀彦（編著）『財務諸表外情報の開示と保証：ナラティブ・リポーティングの保証』（第 7 章）. 同文舘出版.

中村大輔. 2022.「新潟県内上場企業の経営者は COVID-19 の影響をどのように捉えているか：有価証券報告書における MD&A 情報のテキストマイニング分析」『長岡大学研究論叢』20: 205-20.

中村竜哉. 2021.「目標とする経営指標に関する分析：日経平均株価採用銘柄 225 社の有価証券報告書の分析から」『拓殖大学経営経理研究』120: 29–49.

中邨良樹. 2014.「有価証券報告書の記述単語と経営指標との関係に関する一考察」『青山経営論集』49（3）: 102–111.

中山幸雄・津田和彦. 2018.「事業リスクとマネジメントフレームワークに関する考察」『経営情報学会 全国研究発表大会要旨集』13-16.

西野嘉之. 2017.「有価証券報告書の類似度による企業評価」『情報システム学会全国大会論文集』1–6.

野田健太郎. 2016.「有価証券報告書における定性情報の分析と活用：リスクの多様化にともなう望ましい対話のあり方」『経済経営研究』37（1）: 1-51.

廣瀬喜貴・平井裕久・新井康平. 2017.「MD&A 情報の可読性が将来業績に及ぼす影響 : テキストマイニングによる分析」『経営分析研究』33: 87-101.

藤井元雅・坂地泰紀・佐々木一・増山繁. 2020.「有価証券報告書からのリスク文抽出の試み」『人工知能学会第二種研究会資料 2020（FIN-025）』1–5.

藤井元雅・坂地泰紀・佐々木一・増山繁. 2021.「有価証券報告書におけるリスク階層構造分析」『人工知能学会第二種研究会資料 2021（FIN-027）』26-31.

古田成志. 2022.「テキストマイニングを用いたラディカルな組織変革における先行要因の探求：日本企業におけるマトリックス組織の導入状況から」『中京学院大学紀要』1（1）: 1-11.

星野雄介・平尾毅. 2021.「日本企業における「イノベーションという言葉」の普及：有価証券報告書のマイニングを通じて」『武蔵野大学経営研究所紀要』4: 115-42.

矢澤憲一. 2019.「「コーポレート・ガバナンスの状況」の分析：テキストマイニングを利用して」日本会計研究学会第 78 回全国大会.

矢澤憲一・伊藤健顕・金鉉玉. 2021.「テキストマイニングを用いた MD&A，リスク，ガバナンス情報の分析」『青山経営論集』56（1）: 59-84.

矢澤憲一・金鉉玉・伊藤健顕. 2022.「テキストマイニングで解き明かす有報の 60 年」『企業会計』74（2）: 27-34.

吉田慎一郎・中藤哲也・御手洗秀一・廣川佐千男. 2013.「利益伸び率に着目した有価証券報告書のテキストマイニング」『火の国情報シンポジウム論文集 2013（A-5-2）』1–5.

吉田政之. 2018.「リスク情報の可読性と将来業績に関する実証分析」神戸大学大学院経営学研究科大学院生ワーキング・ペーパー.

吉田政之. 2020.「リスク情報開示におけるリスクの種類とその変遷：トピックモデルを用いて」『原価計算研究』44（1）: 116–128.

米田宏生・湯本高行・礒川悌次郎・上浦尚武. 2019.「有価証券報告書の分析に基づく重要な新着ニュースの発見」情報処理学会研究報告（2019-DBS-169 No.19）.

渡部美紀子. 2020.「事業等のリスク情報に関する分析」『人文社会科学論叢』29: 27–43.

渡部美紀子. 2021.「パンデミックに関するディスクロージャーの変化：有価証券報告書の事業等のリスクにおいて説明された内容を中心として」『人文社会科学論叢』30: 61–87.

付録　有価証券報告書に TM を適用した先行研究（五十音順）

文献	サンプル	分析項目	タイプ	分析単位・特徴量・手法など	主な目的と結果
石光(2018)	2004 年度から 2010 年度までの東証一部上場の 3 月決算	・研究開発活動	・予測（仮説検証）	・単語レベル（形態素解析（解析器 MeCab）） ・特徴量(形態素)に名詞，TTR	・研究開発費の金額情報を所与として，TTR の水準が高ければ，将来の ROA の水準も高くなることが示された.
伊藤(2019)	2004 年 3 月から 2014 年 12 月の東証一部上場企業の決算	・MD&A	・記述	・単語レベル（形態素解析） ・特徴量(形態素)に名詞，形容詞 ・分析にトピックモデル	・MD&A の開示の実態を明らかにするためトピックを分析した結果，全体的には MD&A の開示のトピック（資産，事業，利益（費用））は期間中に変化は見られない一方，赤字・減益などの収益状況の違いでは，トピックを通じ開示内容に差が確認された.
伊藤・越智(2016)	2003 年 3 月から 2014 年 12 月の間の継続企業の前提に関する注記を付している非金融上場企業の決算	・継続前提に関する注記	・予測（仮説検証）	・単語レベル（AppleScript の独自コードおよび Casual Conc にて文字数カウント） ・特徴量として注記の文字数の自然対数値 ・分析に回帰分析	・有報において経営者が開示する注記の情報量（文字数）と株主資本コストとの関係の検証した結果，継続企業の前提に関する注記の情報量が多いほど株主資本コストが低下するという仮説は支持されなかった.
延東(2020)	2019 年 4 月時点での機械,電気機器,輸送用機器の 363 社の決算	・事業リスク	・記述	・単語レベル（調達に関するキーワード，調達リスクに関するキーワード（名詞）の抽出）	・3 業種（機械，電気機器，輸送用機器）における調達リスクおよび調達リスクマネジメントの必要性に関する記載動向の違いを有報の事業リスクの項目から確認される.
延東(2022)	2017 年度と 2019 年度の一部上場の輸送用機器産業の企業の決算	・事業リスク	・記述	・単語レベル（調達に関するキーワード，調達リスクに関するキーワード（名詞）の抽出） ・分析にクラスター分析	・2019 年の有報の制度改定が企業の調達リスク開示に与える影響を見るため，改定前後の年度の有報を比較した結果，納期や情報セキュリティなど調達リスクを開示した企業数が増加し，また，QCD や CSR 調達の両リスクを開示するクラスターの企業が増加していることが分かった.
大森(2014)	2007 年度から 2011 年度の 5 年継続している上場している電丸機械製造業の企業の決算	・対処すべき課題	・記述	・単語レベル（形態素解析） ・特徴量(形態素)に名詞 ・分析に共起ネットワーク	・電機業界における各企業の経営課題の認識の違いを，財務業績の変化との関係の解明するため，経営課題と共起しやすい手がかり語（「実施する」など）を元にして，課題を表す単語を抽出し，財務業績の変化のパターンごとに，出現する課題の特徴が示される.

文献	サンプル	分析項目	タイプ	分析単位・特徴量・手法など	主な目的と結果
加藤・五島（2021）	2014 年 5 月から 2019 年 6 月の決算	・MD&A	・予測	・単語レベル（形態素解析（解析器に MeCab）） ・特徴量（形態素）に極性辞書に含まれる品詞，単語のトーン ・極性辞書に Loughran and McDonald（2011）の日本語版 ・分析に回帰分析	・経営者が有報の中で将来見通しを開示する媒体として MD&A の意義を明らかにするため，MD&A のトーンを評価して，トーンが将来業績の予測力を持つことが示された．また，トーンと企業属性および MD&A の章立て変更との交互作用についても検証されている．
北島ほか（2019）	2016 年度の東証第二部上場の製造業の決算	・対処すべき課題	・学習	・単語レベル（形態素解析（解析器に JUMAN）） ・特徴量（形態素）に名詞，TF-IDF	・社是と企業パフォーマンスの関係を検証するため，社是関連の単語が TF-IDF により重み付けられ，2 値分類された企業パフォーマンス（ROA）との関係性について，潜在学習の手法による分類器により良好な結果が得られた．
喜田（2006）	1976 年から 1998 年のアサヒビール（およびキリンビール）の決算	・営業の状況	・記述	・単語レベル（形態素解析（解析器に Text Mining for Celementine（SPSS））） ・特徴量（形態素）に名詞	・組織革新の要因として組織的認知構造の変化を明らかにするため，有報に出現する概念数（名詞の数）に着目し，組織革新との関係性を明らかにした．
城崎ほか（2021）	2019 年の決算（新型コロナウィルス前）	・全ての項目	・記述	・単語レベル（形態素解析） ・特徴量（形態素）に名詞 ・分析に共起ネットワーク	・柔軟な勤務形態を導入している企業の特徴や効果を調査するため，「遠隔」「在宅」「フレックス」の単語を手がかりに，産業，上場市場別の導入傾向の違いや，有報での記載項目箇所の違いがあったことが示された．また，上記の単語の周辺語から，生産性向上，効率化，改革といった業務に対してポジティブな単語が頻出していることが分かった．
Kim and Yasuda（2018）	2002 年と 2003 年の 3 月末決算の上場企業の決算	・事業リスク	・予測（仮説検証）	・単語レベル（リスク関連の単語を抽出） ・特徴量に，リスクの開示量の指標としてそれらカテゴリー数，文字数やセンテンスの数など ・分析に回帰分析	・有報での事業リスクの開示が企業リスクへ及ぼす効果を検証するため，事業リスクから選別したリスクカテゴリーを用いて，事業リスクの開示が企業の総リスクを下げる効果があることを示した．また，事業リスクの開示と企業の総リスクとの間には，正の関係も示され，投資家による企業リスクの評価にも効果があることが示されている．

文献	サンプル	分析項目	タイプ	分析単位・特徴量・手法など	主な目的と結果
金ほか（2022）	2006年から2016年までの一般事業会社の決算	・事業リスク ・対処すべき課題 ・MD&A	・予測（仮説検証）	・単語レベル（形態素解析（解析器にMeCab，辞書にipadic, UniDic, Neologd，及びユーザー辞書）） ・特徴量に可読性，単語のトーン ・分析に回帰分析	・経営者交代が（財務情報以外の）記述情報の変化に与える影響を検証するため，有報の対象項目の記載内容から，交代により過年度の記述情報が再利用される度合い（スティッキネス）は低下し，可読性が向上，さらに，記述情報のトーンがポジティブになることが明らかにされた．
許・金（2021）	2008年から2010年までのリーマンショックに倒産した上場企業54社（倒産直前の年と2年前の2年分を利用）とその後10年以上継続している上場企業87社	・提出会社の状況（配当政策）	・予測	・単語レベル（形態素解析（解析器にMeCabのほか，辞書にipadic，複合語については専門用語自動抽出システム（TermExtract）を使用）） ・特徴量（形態素）に名詞，動詞，形容詞，TF-IDF ・分析に共起ネットワーク（係り受けが考慮され，解析にはCaboChaが利用） ・分析に回帰分析	・財務データと有報のテキストデータとを結合し，企業の倒産の判別するために，2年間分のデータを用い財務データと特徴量を組み合わせることで，企業の倒産の判別の精度が高まった．
小寺ほか（2018）	2008年から2016年の一部上場企業の決算	・対処すべき課題	・学習	・センテンスレベル（係り受け解析にCaboCha） ・特徴量（形態素）に課題文識別のための文末表現 ・分析に判別モデル（機械学習）	・対処すべき課題の項目から，（因果関係文とは異なる）企業の課題文の抽出するため，107個の手がかり表現により構築された判別モデルにより，課題文が良好な精度で抽出された．
小寺ほか（2019）	2008年から2018年の決算	・対処すべき課題	・学習	・センテンスレベル（係り受け分析にCaboCha） ・特徴量に文末表現 ・分析に判別モデル（機械学習）	・未来表現文は時制の判別により，また目的手段文は係り受け解析による判別モデルにより，未来表現文と目的文の特徴を兼ね備えた未来志向文の抽出について良好な結果が得られた．
近藤・石光（2016）	2004年から2016年の上場企業の3月決算	・事業リスク ・ガバナンス	・記述 ・予測（仮説検証）	・単語レベル（形態素解析（解析器にMeCab）） ・特徴量（形態素）に（内部統制に関連する）名詞，TF-IDF，TTR ・分析に回帰分析	・事業リスクおよびガバナンスに関する記載状況の把握のため，事業リスクおよびガバナンスの記載内容に関する特徴量の推移を示し，さらに，事業リスクより定量化した不確実性とガバナンス項目より抽出した内部統制の相関関係および内部統制による企業の将来業績への影響について検証される．

文献	サンプル	分析項目	タイプ	分析単位・特徴量・手法など	主な目的と結果
近藤・石光（2020）	近藤・石光（2016）と同様	・ガバナンス	・予測	・単語レベル（形態素解析（解析器に MeCab）） ・特徴量（形態素）に（内部統制に関連する）名詞，TF-IDF ・単語間の類似度の測定に単語分散表現（Word2Vec）が利用 ・分析に回帰分析	・MC と将来業績との関係を検証するため，有報のガバナンスの記載内容から内部統制に関するキーワードとその類似語より内部統制の実施を定量化し将来業績と有意な関係を検証し，将来業績への部分的な影響が示される．
佐久間・田中（2018）	TOPIX1000 銘柄の 2008 年 3 月期から 2016 年 3 月期の決算	・業績等の概要	・記述	・センテンスレベル（因果関係文を抽出） ・意味解析（因果文の中の単語の極性）	・投資判断に重要な有報の因果関係文の極性（トーン）を調査するため，各極性の企業数の推移，極性と業績との関係性，さらに，極性と 1 期先の将来業績との関係が検証された．
佐藤ほか（2018）	2008 年から 2016 年までの TOPIX 1000 の企業の決算	・業績等の概要 ・対処すべき ・事業リスク	・学習	・センテンスレベル（係り受け分析に Cabocha，手がかり語などの素性作成のため形態素解析に MeCab） ・特徴量（形態素）に助詞のほか手がかり表現 ・分析に判別モデル（SVM）	・有報の記載から手がかり語を使って投資判断に有益な因果関係文が良好な精度で抽出するための判別モデルが提案された．
佐藤（隆）ほか（2021）	2014 年 1 月から 2017 年 12 月の東証一部上場企業の決算	・事業リスク	・予測（仮説検証）	・単語レベル（形態素解析（解析器に MeCab）） ・特徴量（形態素）に名詞 ・単語間の類似度の測定に単語分散表現（Word2Vec）が利用 ・分析に回帰分析	・事業リスクのテキストデータは，以下の発見より，株価のリスクと整合的であることが示される．つまり，（1）独自の辞書（Campbell et al.（2014）などをもとに作成）により，事業リスクにおけるリスク関連の単語と株式収益率から推定されるリスク指標との間に正の相関，（2）リスクの単語の出現確率と 3 ファクターローディングについて正の相関，があった．
佐藤（慧）ほか（2021）	日経 225 銘柄の決算（年代不明）	・全ての項目	・学習	・センテンスレベル ・特徴量（形態素）に名詞 ・形態素解析をベースに文と文の単語の類似度の測定には単語分散表現（Word2Vec）が部分的に利用 ・分析に深層学習	・有報における記載の中から企業理念文および業績要因文（高野（2019）に依拠）を深層学習にて自動抽出をできるよう学習モデルを構築し，さらに，企業理念と業績要因の関連性についても良好な精度で推定できた．

文献	サンプル	分析項目	タイプ	分析単位・特徴量・手法など	主な目的と結果
首藤・緒方（2009）	2006年の決算（大企業，中小企業及び新規上場企業に分類）（2007年7月25日以前に提出された有価証券報告書のうち，直近に提出されたものを集計）	・MD&A	・記述	・その他（MD&Aの掲載社数，表・グラフの記載数，表・グラフの記載内容，他の記載項目の参照など）	・MD&Aの記載内容それぞれで用いられる記載要素（表，グラフ，他の記載項目参照）が集計され，新規上場企業がMD&Aの開示には積極的であった．また，MD&Aでは有報における他の項目を参照しているのが分かった．
白田・坂上（2008）	・2000年から2003年に倒産した企業と同年に継続していた企業の決算	・全体の項目（注記・特記情報を含む）	・記述	・単語レベル（形態素解析（解析器にChaSen）） ・特徴量（形態素）に名詞，動詞，形容詞，助詞，TF-IDF	・2000年4月以降にわが国において破綻した上場企業の倒産の直前期と同時期の継続企業の有報に出現する単語の傾向を比較することで倒産の兆候が確認された．
白田ほか（2009）	・1999年から2005年に倒産した上場企業（90社）の倒産直前期の決算，および2005年に現存していた上場企業（90社）の決算	・全ての項目	・記述	・単語レベル（形態素解析にOmniFind（IBM）） ・特徴量（形態素）に名詞，動詞，形容詞，形容動詞	・倒産企業の特徴を明らかにするため，倒産群と継続群それぞれの有報に出現する特徴語（名詞・動詞・形容詞・形容動詞）を集計し，双方の群でそれぞれ条件付けられた特徴語の出現確率に差が見られた．
鈴木ほか（2021）	chABSA-dataset*	—	・学習	・センテンスレベル ・分析に学習言語モデル（BERTなど）	・有報を含む日本語金融コーパスから，金融分野の事前学習言語モデルを構築し（BERTモデルやELECTRAモデル），金融分野のテキストデータに関するタスク（因果関係文の抽出やトーン分析など）を実験することで，汎用的なモデルよりも高い性能が得られた．
鈴木ほか（2022）	2018年2月から2020年12月の決算（その他決算短信およびWikipedia）	—	・学習	・単語レベル（形態素解析（解析器にMeCab，辞書にipadic）） ・特徴量に可読性（そのための品詞に漢語，和語，動詞，助詞が），トーン	・有報を含む金融テキストとWikipediaからのコーパスを利用して，（追加）事前学習を行い，金融用語に特化した幾つかの金融言語学習モデルが構築され，評価用のタスク（chABSA-datasetを用いた極性評価）が遂行され，その結果，BERTモデルが最も精度が高かったとの結果が得られている．

文献	サンプル	分析項目	タイプ	分析単位・特徴量・手法など	主な目的と結果
陶（2022）	「記述情報の開示の好事例集 2020」（金融庁）の中から 30 社の決算	・事業の状況 ・経理の状況	・記述 ・予測（仮説検証）	・単語レベル ・特徴量に文字数 ・分析に相関分析，線形判別分析	・上場企業のリスク情報の開示の状況を明らかにするため，コロナ禍で影響を受けた業種の中の企業の業績指標や企業属性（規模など）とコロナ関連の記述量（文字数）との間に関連性が認められ，また，コロナの影響で業績悪化した業種の企業では，業績改善した業種と比較して，コロナ関連の開示文字量が増加する傾向にあったことが明らかにされた．
高野ほか（2022）	2012 年 1 月から 2021 年 12 月の間の決算企業から 160 銘柄を無作為抽出（学習データは，2012 年 1 月から 2018 年 12 月の 100 銘柄で，テストデータはこれと異なる銘柄）	・提出会社の状況（配当政策）	・学習	・単語レベル（形態素解析（解析器に MeCab）） ・特徴量（形態素）に名詞，動詞，形容詞，副詞 ・分析にトピックモデルおよび学習言語モデル（BERT）	・配当政策の文を自動抽出するため，トピックモデルをもとに将来の配当政策文を正例とする学習データを用意し，BERT による自動抽出モデルが最も良好な結果であった．
高野ほか（2019）	2013 年から 2018 年 5 月の間の上場企業の決算	・全体の項目	・学習	・単語レベル（形態素解析（解析器に MeCab）） ・センテンスレベル（係り受け解析に CaboCha） ・特徴量（形態素）に名詞，手がかり語 ・分析に深層学習	・有報から企業の事業セグメント名の抽出と，事業セグメントごとの業績要因文および業績結果の自動抽出方法の提案するため，手がかり語より有報から業績要因文を抜き出す一方，事業セグメント名を「企業の概要」の中の「従業員の状況」から抽出し，両者を結びつけるため深層学習を通じて，良好な評価が得られた．
竹内ほか（2008）	1999 年から 2005 年の間の倒産した上場企業 90 社の決算，継続企業には 2005 年の決算	・提出会社の状況（配当政策）	・記述 ・予測	・単語レベル（形態素解析（解析器に OmniFind（IBM））） ・センテンスレベル（OmniFind（IBM）） ・特徴量に文脈語を設定	・倒産企業の特徴の解明するため，継続群と倒産群の 2 群を有報における配当政策の記載内容について比較し，倒産群を特徴づける表現を単語・文脈レベルで抽出し，継続群には内部留保に関する特徴的記載が見つかる一方，倒産企業の将来予測には結びつかなかった．
辻野ほか（2010）	・2008 年度の上場企業の決算 ・上記以外の 2004 年度から 2007 年度の上場企業の決算	・対処すべき課題	・記述	・単語レベル（形態素解析） ・特徴量（形態素）に名詞，TF-IDF ・分析にトピックモデル（潜在意味解析）	・経営環境の悪化した 2008 年度において特徴ある経営課題と対策を抽出するため，TF-IDF により対処すべき課題を定量化し，さらに，潜在意味解析によるクラスタリングにより，4 つの企業群が識別され，それぞれ主要な経営課題の特徴が示される．

文献	サンプル	分析項目	タイプ	分析単位･特徴量･手法など	主な目的と結果
土屋(2018)	2016 年 4 月から 2017 年 3 月の間の上場企業の決算	・事業リスク	・記述	・単語レベル（形態素解析） ・特徴量(形態素)に名詞，TF-IDF ・分析にクラスター分類	・事業リスクの項目も類似性について，いくつかの業種で類似性が確認され，事業リスクの記載内容の多様性は業種ごとに差があることが示される.
土屋(2020)	2018 年 4 月から 2019 年 3 月までの上場企業の決算（かつ 2019 年 3 月末で東証一部銘柄）	・事業リスク	・記述	・単語レベル（形態素解析） ・特徴量(形態素)に名詞 ・分析にトピックモデル	・事業等のリスクの記載内容の類似性について，22 業種について業種ごとの 30 トピックの記載割合が明らかになり，異業種であっても，認識しているリスクに類似性が確認される.
土橋・中田(2021)	2017 年 と 2018 年において上場企業からランダムに抽出した企業 70 社の決算	・対処すべき課題 ・事業リスク	・学習	・単語レベル（形態素解析（解析器に MeCab)） ・センテンスレベル ・分析に学習言語モデル（BERT）	・有報から ESG に関する記述を自動抽出するため，ESG 関連文の学習データセットを作成し,(対象文を分かち書きにより単語レベルから予測するモデルに比べて) BERT（ファインチューニングあり）による学習モデルにより，年度・企業が異なるデータに対して精度良く汎化し, ESG 関連文の自動抽出が可能になる.
中島(2020)	2014 年 1 月から 2018 年 12 月の間の上場企業の決算	・業績等の概要 ・対処すべき課題 ・事業リスク ・ＭＤ＆Ａ ・ガバナンス ・注記事項	・予測	・単語レベル（形態素解析（解析器に MeCab，辞書に ｉｐａｄｉｃ-NEologd)） ・特徴量(形態素)に記号意外全ての品詞，TTR，可読性 ・分析に回帰分析	・有報の非財務データの特徴量の決定要因を探索するため，各特徴量（文字数，TTR，可読性）を被説明変数とし，企業特性を説明変数とする回帰モデルの検証から，減益企業よりも損失企業の方が可読性が低下する結果が示される.
中野(2010)	2002 年から 2007 年の間の上場企業（金融業除く）の 3 月決算	・事業リスク ・MD&A	・記述 ・予測	・単語レベル ・特徴量に開示量（文字数および段落数） ・分析に回帰分析	・有報における非財務情報の開示行動の実態および開示行動への影響要因の探索から，事業リスクと MD&A の文字数との間には相関が確認さら，事業リスクおよび MD&A の双方の開示量は，企業規模や市場からの注目度（アナリストカバレッジ）が影響してることが示される.
中村(2022)	2018 年度から 2020 年度の新潟県内の上場企業 35 社の決算	・MD&A	・記述	・単語レベル（形態素解析（解析器に ChaSen)） ・特徴量(形態素)に名詞 ・分析に共起ネットワーク	・COVID-19 の財務的な影響を新潟県内上場企業の経営者がどのように分析しているかを調査する中で，MD&A の記載量は増加傾向にあり，COVID-19 関連の単語も増加し，さらに，それによる企業業績へのネガティブな影響も確認している.

文献	サンプル	分析項目	タイプ	分析単位・特徴量・手法など	主な目的と結果
中村(2021)	2020 年度（2021 月 4 月末時点）の日経銘柄 225 社の企業	・経営方針 ・MD&A ・事業リスク	・記述	・単語レベル ・特徴量に各種経営指標の単語	・有報の各項目において経営指標（ROE，資本コスト，ESG 要素，SDGs，気候変動など）の記載状況を明らかにする中で，特に，リスク項目に感染症の記載する企業の多さが顕著だったことがわかる．
中邨(2014)	2011 年度から 2012 年度の上場企業 5 業種の 296 社の決算	・全ての項目	・記述 ・予測	・単語レベル（形態素解析（解析器に MeCab）） ・特徴量(形態素)に名詞 ・方法に因子分析，回帰分析	・企業活動とその経営指標との関係性の解明のため，有報のテキストデータの中の出現率上位の単語から，企業活動に関する「資金性」と「時間性」の因子を抽出し，それら因子を構成する単語と経営指標との関係が検証されている．
中山・津田(2018)	2013 年から 2017 年の 3 月決算	・事業リスク	・記述	・単語レベル ・特徴量に文脈文（ISO マネジメント関連）の中の名詞	・ISO マネジメントの実施状況を明らかにするため，有報の記載の中に「ISO」を含む企業を対象にしても，「品質」が事業リスクの項目に出現する企業と出現しない企業があり，産業によっても傾向が異なるなど，ISO マネジメントの実施のばらつきが示される．
西野(2017)	2014 年から 2017 年の対処すべき課題の記載内容の変化のない東証一部上場企業 36 社の決算	・対処すべき課題	・予測	・単語レベル（形態素解析） ・特徴量に（対処すべき課題の記載の前年度との）類似度	・企業内の有報の変化を比較するため，対処すべき課題の類似度を数値化し，同一企業の財務データに対応付ける方法が提案され，対処すべき課題の内容が全く変わらない企業が存在し，そうした企業は概ね財務成果（売上高および当期純利益）が安定していることが判明した．
野田(2016)	2003 年から 2012 年の東証一部上場企業（金融等除く）の決算	・対処すべき課題 ・事業リスク ・MD&A ・ガバナンス	・記述 ・予測	・単語レベル（形態素解析） ・特徴量(形態素)に記述量，特に名詞 ・方法に回帰分析	・有報の各項目での開示量（記述量）の変化を明らかにし，さらに，開示量に影響を及ぼす特性（中野(2010）に依拠）を発見している．さらに，MD&A の開示は市場から好評価といった開示による投資家からの評価が検証される．
廣瀬ほか(2017)	2004 年から 2015 年の間の上場企業の 3 月期決算	・MD&A	・予測（仮説検定）	・単語レベル（形態素解析（解析器に MeCab）） ・特徴量に可読性（難易度，述語数，平仮名率，文字数など） ・方法に回帰分析	・MD&A の文書の可読性（総文字数および難易度）の決定要因の解明および可読性が利益の持続性に与える影響の検証において，特定の企業特性で MD&A の可読性が低下する傾向で，また，可読性（総文字数)は 1 期先の業績(ROA)に負の影響が示された．

文献	サンプル	分析項目	タイプ	分析単位・特徴量・手法など	主な目的と結果
藤井ほか（2020）	日経225社からランダム抽出した70社の決算	・事業リスク	・学習	・センテンスレベル（係り受け解析器 に CaboCha）（ただし，学習の際の素性を与える上で形態素解析（MeCab）を利用） ・分析に学習言語モデル（BERT）	・有報の中で企業が認知するリスク関連の文を自動抽出するための分類器の開発のため，事業リスク項目の中のリスク文を仮定表現の文と可能性表現の文に分類した上で，どの分類器手法が有効かを調べるために評価実験の結果，BERTの学習（ほかに，ロジスティック回帰，ランダムフォレスト，SVM）の精度が高かった．
藤井ほか（2021）	2018年度の日経225社からランダムに抽出した71社の決算（PDFファイルをテキスト化）	・事業リスク	・記述	・単語レベル（人手によりリスク関連の因果関係文からリスク要因の単語を抽出）	・事業リスクからリスク関連の文を抽出し，因果文の関係から，リスク要因を特定し，原因と結果に識別することから，リスク要因から業績・経営成績までのリスク階層構造が示され，企業価値評価への可能性が検討されている．
古田(2022)	2004年から2020年の東証一部企業の決算（マトリックス組織を導入した51社の前年度の決算）	・対処すべき課題	・記述	・単語レベル（形態素解析（分析器に Text Mining Studio（NTT DATA）） ・特徴量(形態素)に名詞・動詞含む全品詞 ・分析に共起ネットワーク	・マトリックス組織導入の先行要因を明らかにするため，導入前年度の有報から単語頻度および共起ネットワークより探索と活用に関した二重性の傾向が確認された．
星野・平尾（2021）	2004年から2020年の間の上場企業の決算（四半期報告書，半期報告書も含む）	・全ての項目	・記述	・単語レベル（形態素解析（解析器に MeCab）） ・特徴量(形態素)に名詞，他一部動詞を除いた品詞 ・分析に共起ネットワーク	・有報の中でのイノベーションに関する単語の普及の調査において，イノベーションに言及する企業数や文書数が増加傾向の一方，産業ごとでその言及の程度に違いのあることが明らかになる．また，イノベーションは，ポジティブな意味を持つ単語と共起することも分かった．
矢澤(2019)	2018年3月期から金融業に属する企業の決算	・ガバナンス状況	・記述 ・予測	・単語レベル（形態素解析） ・特徴量に（コーポレート_ガバナンスの質の評価指標に関する）名詞 ・方法に回帰分析	・コーポレート・ガバナンスの質の評価指標（CGAI）に従い，有報のガバナンス状況の項目の記載字数，インプット（独立性，客観性など），プロセス（計画，実行，報告など）への説明変数が検証される．

文献	サンプル	分析項目	タイプ	分析単位・特徴量・手法など	主な目的と結果
矢澤ほか(2021)	2004 年から 2018 年の全上場企業（金融業は除く）の決算	・MD&A ・事業リスク ・ガバナンス状況	・記述 ・予測	・単語レベル（形態素解析（解析器に MeCab，辞書に ipadic, UniDic, Neologd など）） ・ユーザー辞書には KH Coder から抽出した複合語を利用 ・特徴量(形態素)に文字数，数字・固有表現, 単語数, 可読性，センチメントなど ・方法に回帰分析	・有報の各項目においてテキストの特徴量（定量化指標）の推移を明らかにした上で，各項目の特徴量に影響する企業属性（ファンダメンタル要因）を探索し，MD&A や事業リスクのテキストデータの量やその質（具体性や可読性）に業績や財政状況が影響しているのが示された.
矢澤ほか(2022)	1961 年から 2020 年の間の企業の決算	・全ての項目	・記述	・単語レベル（形態素解析（解析器に MeCab）） ・形態素解析に独自の辞書を利用 ・特徴量(形態素)として一般名詞の他に形容詞 ・トーン（センチメント）分析	・長期に及ぶ有報の記載内容の分析を行うため，1961 年から 2020 年に及ぶ 60 年間の記載量（頁数）の推移に加えて，2004 年から 2020 年の間の有報の項目毎の記載量（単語数）の推移，さらに，事業の状況における特定キーワード（財政状態，経営成績，リスク）の頻度が明らかにされ，さらに，事業の状況で出現する単語のトーンも経済危機や業種によって異なっていた.
吉田ほか(2013)	2006 年から 2012 年の医薬品関連企業の決算企業	・企業の概況および事業の内容の項目	・探索	・単語レベル（形態素解析（抽出には，汎用連想計算器（GETA）が利用）） ・特徴量(形態素)に名詞および動詞	・有報の中の特徴語により利益が継続して増える企業を特定するため，重要でかつ多くの有報の項目で出現する特徴語およびそれら組み合わせで検証されている.
吉田(2018)	2004 年から 2017 年の 3 月の上場企業（金融業除く）の決算	・事業リスク	・予測（仮説検証）	・単語レベル（形態素解析（解析器に MeCab）） ・特徴量に可読性（文字数と難易度） ・方法に回帰分析	・リスクマネジメントの質による将来業績予測のため，質を可読性（文字数と難易度）で評価することで，リスクの記述量は将来業績（ROA）に負に影響する一方，文書の難易度は，直接の有意な効果はないことが判明した（ただし，可読性の指標は当期の業種調整 ROA との交互作用効果が確認).
吉田(2020)	2004 年から 2017 年の 3 月の上場企業（金融業除く）の決算	・事業リスク	・記述	・単語レベル（形態素解析） ・特徴量に名詞 ・分析にトピックモデル	・有報で認識されるリスクの種類とその変遷を捉えるため，事業リスクの記載内容からトピックが識別される.

文献	サンプル	分析項目	タイプ	分析単位・特徴量・手法など	主な目的と結果
米田ほか（2019）	chABSA-dataset* のほか，新聞経済面に掲載された記事（2014年1月から3月））	—	・学習	・センテンスレベル（因果文を抽出） ・因果関係文から手がかり語を頼りに名詞（JUMANにより形態素解析）で構成される経済的影響のある事象間のパターンの抽出	・ニュース記事（新聞）から経済的に影響のある事象を見つけるため，有報のテキストから経済的影響のある因果関係文から抽出した事象パターン（いくつかの名詞の組み合わせ）をもとに経済事象を抽出し，ニュース記事から経済的に影響のある事象を自動抽出することが試みられた．
渡部（2020）	2017年から2018年の3月の上場企業13社の決算	・事業リスク（統合報告書も併用）	・記述	・単語レベル ・リスクとして説明されている項目を用語が抽出 ・分析に数量化3類	・金融庁による記述情報の開示の好事例集にみるリスク項目開示の動向を調査し，事業リスクの関連の用語で企業分類する2軸（概ね自然・環境・社会の軸と経済・法律の軸）が識別され，さらに，有報と統合報告書におけるリスクについての認識の違いが示された（有報ではリスクについて網羅性が重視）．
渡部（2021）	2019年3月時点で有報の事業リスクでパンデミックを言及していた企業の2019年と2020年3月決算	・事業リスク	・記述	・単語レベル ・文字数，パンデミックについての対処や影響について記述されている用語が抽出 ・分析に数量化3類	・パンデミック前後での有報の記載内容の変化を見るため，パンデミック前に既に言及していた企業では，パンデミック後にパンデミックを表す単語およびその対処や影響に関する記載が増加する傾向であり，さらに，パンデミック前後で，用語を用いた数量化3類の分析で識別された2軸によるに変化が見られた（パンデミック後ではパンデミックの軸が確認）．

* 2016年の有報内のセンテンスに極性が与えられたデータセットである（https://github.com/chakki-works/chABSA-dataset）．

Text Mining for Text Data of Annual Securities Reports: A Literature Review

Takahito KONDO
Yu ISHIMITSU

ABSTRACT

In this paper, we provide a literature review for discussing an application of text mining to the analysis of text data of annual securities reports submitted by Japanese companies, show implications for the application, and suggest directions for future research in accounting.