機械学習を用いた統合報告書の ESG 関連ページの推定

Estimation of ESG-related pages in integrated reports using machine learning

河村 康平[1]　高野 海斗[2]　酒井 浩之[1]

永並 健吾[1]　中川 慧[2]

Kohei Kawamura[1], Kaito Takano[2], Hiroyuki Sakai[1]

Kengo Enami[1], Kei Nakagawa[2]

[1] 成蹊大学
[1] SEIKEI University
[2] 野村アセットマネジメント株式会社
[2] Nomura Asset Management Co., Ltd.

Abstract: 近年の資産運用分野では，財務情報である売上や利益だけでなく，非財務情報である環境(Environment)，社会(Social)，企業統治(Governance)の 3 つの観点を考慮して投資を行う「ESG 投資」が世界的に広まりつつある．そのため ESG 投資において，企業による ESG 情報を判断材料として獲得することは重要である．日本においては，企業の自社の ESG 情報を開示する手段として統合報告書がある．決算短信や有価証券報告書のような他の金融テキストとは異なり，統合報告書には企業の財務情報に加え非財務情報が開示されており，ESG 投資を行う上で重要な情報源であると言える．しかし，統合報告書の中には 100 ページを超えるものもあり，人手で ESG 情報を探すには多くの時間と労力が必要となる．統合報告書において ESG 情報が存在する位置を自動で推定することが可能になれば，例えば，企業の成長・持続可能性を判断し，投資リスクを抑えることが可能となる，そこで本研究では，ESG に関連する内容についての記述が含まれるページを ESG 関連ページと定義し，機械学習手法を用いて統合報告書から ESG 関連ページを推定する手法を提案する．

1. はじめに

近年資産運用の分野では，財務情報である売上や利益だけでなく，非財務情報である環境(Environment)，社会(Social)，企業統治(Governance)に対する企業の取り組みを考慮して投資を行う「ESG 投資」が世界的に広まりつつある[1]．これは，短期的な利益追求による金融危機の反動から企業の長期的な価値が重要視されるようになり，目の前の売上や利益にとらわれない長期的・持続的な展望への関心が高まっていることが背景にある．

資産運用分野での世界的な ESG 投資に対する関心の高まりを受け，投資家など様々なステークホルダーは ESG 情報などの非財務情報の開示を企業に求めるようになり，企業の ESG 情報(環境，社会，企業統治の問題に対する取り組みについての情報)の開示の重要性は高まり続けている．日本においても，ESG 情報の開示企業は，近年増加傾向にある[2].

一方で，機関投資家や個人投資家にとって企業が開示している ESG 情報の獲得・利用には 2 つの問題が存在する．1 つは企業の ESG 情報の開示媒体が多様なことにより ESG 情報の満足な活用が出来ていないことが挙げられる．企業は ESG 情報の開示手段として統合報告書，CSR 報告書，有価証券報告書，自社ホームページなど多くの媒体を使用しており，全ての内容を確認するには多くの労力と時間が必要となる．2 つ目は，比較可能性の欠如である．統合報告書，CSR 報告書，サステナビリティレポートなどの任意開示は ESG 情報のような非財務情報を多く含む一方で，財務情報のように開示に関する統一的な基準が存在しないために，情報開示の量と質共に各企業の裁量に委ねられることが多い，そのため．同業他社，あるいは同企業の年ごとの比較や分析などが容易でなく，投資家の投資判断材料とするには困難である．

上記の 2 つの問題を解決するためには，企業が公

開している媒体から ESG 情報を自動で抽出し，ESG 情報の比較や定量的分析が可能な形式に変換できる仕組みが必要である．我々は，このような仕組みの実現を目指し，第一段階として，企業の ESG 情報が掲載されている媒体として統合報告書に注目し，統合報告書の ESG 情報が含まれる可能性の高いページを，機械学習モデルにより自動で推定することを試みる．

統合報告書とは，売上や利益などの財務情報に加え，ESG 情報などの非財務情報が開示された報告書である．[3]の調査では，企業への「ESG への取り組みについて情報を開示している媒体をお答えください（複数選択可）」という問に対して，解答数 500 のち 50.2%の企業が「統合報告書」と解答している．これは，ホームページと IR 説明会資料を除いた中では最も高い結果であり，統合報告書は企業が発行する報告書の中では ESG 情報を獲得しやすい媒体であると言える．しかし統合報告書に対して機械学習や統計的分析を適用することは今まで困難とされてきた．その理由は，統合報告書は任意開示であり，開示に関する統一的な基準が存在せず，開示内容が企業毎に大きく異なるからである．また，開示形式についても，決算短信や有価証券報告書のような他の金融テキストのように共通した形式が存在せず，企業毎に独自のデザインや図表を用いており，PDF ファイルをテキストデータに変換する際にレイアウトが崩れてしまうなどの問題もあった．

本研究で扱うデータは，統合報告書の PDF ファイルをテキスト情報に変換したデータである．そのため，先述したように文の前後が変わる，文が分割される，文字化けが起こるなど元のレイアウトの情報を損なうこともある．提案手法では，このような文やレイアウトが崩れたテキストデータに対して，自動生成した学習データを用いて機械学習モデルを学習し，統合報告書のページ単位でマルチラベル分類を行うことで，E,S,G それぞれについて関連性の推定を試みる．さらに，本研究で扱うような形式が崩れたテキストデータに対して，機械学習モデルの適用がどの程度可能なのかを検討する．

2. 関連研究

本研究の特徴は以下の通りである．

（1）対象は統合報告書
（2）学習データの自動生成
（3）ページ単位でのマルチラベル分類

これらの点を踏まえて関連研究を挙げていく．

ESG 情報の抽出に関連する研究として[4]がある．[4]は有価証券報告書から ESG 関連文（ESG に関連する文）を抽出する手法を提案した．[4]と本研究の違いは対象が統合報告書であることと，ESG 情報の抽出をページ単位で行うことである．統合報告書から ESG 情報を文単位で取り出すことは可読性や分析のしやすさから需要があるが，統合報告書 PDF のテキスト化の際に文の形が崩れてしまうことが多いことから，本研究ではページ単位で抽出することにした．また，統合報告書には有価証券報告書などの金融テキストに比べ ESG 情報が豊富に含まれるため文として抽出するメリットは少なく，統合報告書はページごとに内容がある程度まとまっている点からも，統合報告書においてはページ単位での ESG 情報の抽出が有効であると考える．

統合報告書を用いたテキストマイニング分析に関連する研究として[5]がある．[5]では，環境報告書やサステナビリティ報告書，統合報告書の経営トップメッセージのテキストに対して，深層学習により自動的に社会，及び環境ラベルを付与することで，3 つの媒体の違いによる環境および社会情報開示の記述量の傾向を分析した．[5]は，テキストマイニングの分析対象を統合報告書の中の経営トップメッセージのみに絞り込んでいるが，統合報告書には，ESG 情報が経営トップメッセージ以外の場所にも豊富に掲載されているため，より広範囲の情報を利用できた方が投資において良い判断材料になるはずである．我々は，統合報告書の一部ではなく，全てのページを機械学習の適用範囲とするため，統合報告書の広範囲の内容から ESG 情報を獲得することができる．また，本手法により抽出した ESG 関連ページを用いることで，統合報告書の広範囲の内容を対象としたテキストマイニング分析が可能になると考える．

金融テキストに対してページ単位で機械学習モデルを適用した研究として，[6]がある．[6]では株主招集通知という金融テキストから，ルールベースで自動生成した学習データを用いて深層学習モデルを学習し，株主招集通知から自動的に重要ページを抽出する手法を提案した．[6]では本研究のように形式が崩れたテキストデータを扱っており，このようなデータに対して様々なモデルの提案や考察を行った．[6]と本研究は，単語の出現情報を利用したルールベースにより学習データを自動生成する点で共通しているが，本研究の処理対象は統合報告書であるため，学習データの生成に使用する単語が異なる．ESG 関連ページには必ず出現するような単語は存在しないが，高い確率で出現する単語は存在するため，我々はこのような単語を利用し学習データを自動生成する．また，ページに対して単一ラベルではなくマルチラベルを付与する点でも異なる．

3. 提案手法

統合報告書は，1 つのページに ESG の複数のトピックの内容が開示されることもある．本研究では，統合報告書のあるページについて，そのページが E,S,G のそれぞれのトピックに関連するかどうかをマルチラベル分類により推定する手法を提案する．

ただし ESG については，一般にどのような話題が該当するのか明確に定められておらず，投資家によって解釈が異なる．本研究では，表 1 に示す話題を ESG と定義する．また，図 1 に ESG 関連ページの例を示す．

表 1. 本研究における ESG の定義

トピック	話題
E	気候変動，生物多様性，汚染， 資源利用，有害物質の排出と廃棄物
S	ダイバーシティと機会均等， 地域社会，人材育成， 顧客の健康と安全に対する責任， 非差別，労働基準・雇用
G	企業倫理，反競争的行為， 汚職・腐敗の防止， リスクマネジメント， コーポレート・ガバナンス， ステークホルダー・エンゲージメント

図 1. ESG 関連ページの例

3. 1. 提案手法の概要

提案手法の概要を以下に示す．

Step1: 統合報告書 PDF ファイルをテキストデータに変換し，さらにルールベースによりテキストデータを整形する．

Step2: Step1 で取得した統合報告書のテキストデータを用いて E,S,G それぞれに対する特徴語を抽出する．

Step3: Step1 で収集したページに対して，特徴語を用いたルールベースによりラベルを付与し，学習データを自動生成する．

Step4: 学習したモデルを用いて，統合報告書から ESG 関連ページを推定する．

3. 2. 統合報告書 PDF のテキスト抽出

統合報告書の各ページを機械学習モデルに入力するために，まず PDF ファイルをテキストデータに変換する必要がある．PDF ファイルをテキストデータに変換するために，MacOS に標準インストールされている「Automator」を使用した．「Automator」を用いることで，PDF ファイルからテキストデータをページごとに取得することが可能である．また．他のソフトに比べて比較的段組みを考慮して PDF ファイルをテキストデータに変換出来る．しかし「Automator」による変換により，不要な空欄や改行が挿入されるため，処理がしやすいように可能な限りきれいな形に整えるために，ルールベースによるテキストデータの整形を行う．具体的には以下のようなルールを設ける．

・数値を全て 0 に変換する．
・空欄を空文字に変換する．
・句点”。”を改行に変換する．
・品詞が助詞・助動詞で終わった場合，後ろの文と連結する．など

4 つめに示した文の連結を行うルールは，本来繋がりのない文が後ろにあった場合は適切に文を連結出来ないが，このような場合よりも，連結ルールにより正しく文を修正できた場合の方が多かったため，ある程度の誤りは許容することにした．

3. 3. 特徴語の抽出

特徴語は以下の手法で獲得される．

Step1: 統合報告書から単語“ESG”を含むページを取得し，取得したページの文集合から Word2Vec により 100 次元の分散表現を学習する．

Step2: 統合報告書に出現する名詞nと，$w \in${“環境”，“社会”，“ガバナンス”}に対して正規化自己相互情報量$npmi(w,n)$を計算する．

Step3: ある企業tの統合報告書に名詞nが出現する確率$P(w,n)$に基づくエントロピー$e(n)$を計算する．

Step4: Step2，Step3 で計算した$npmi(w,n)$と$e(n)$が共に上位の名詞nを特徴語候補とする．

Step5: Step1 で学習した 100 次元ベクトル空間上で，特徴語候補をk-means 法($k=15$)でクラスタリングする．

Step6: Step6 でクラスタリングされたクラスタから E, S, G ごとに対応するクラスタを選択し，それらの中の名詞を人手で選別することで，各クラスタ内の名詞を特徴語として獲得する．

Step2 において，環境，社会，ガバナンスのそれぞれと多く共起する単語は環境，社会，ガバナンスのそれぞれと関連性が高いという考えに基づき，$w \in${“環境”，“社会”，“ガバナンス”}と，統合報告書に出現する名詞nについて式 1 により自己相互情報量を求める．なお，本研究ではある 2 つの単語が同じページに出現することを共起とする．

$$npmi(w,n) = \frac{pmi(w,n)}{-\log_2 P(w,n)} \quad (1)$$

$$pmi(w,n) = \log_2 \frac{P(w,n)}{P(w)P(n)} \quad (2)$$

$P(x)$: 統合報告書のページ集合において単語xが出現する確率

$P(x,y)$: 統合報告書のページ集合において単語x,yが同時に出現する確率

式(1)は式(2)の値を正規化したものであり，x,yの希少度による影響を軽減し，ノイズとなる単語（人名など）を除くことができる．

Step3 における，ある企業tの統合報告書に名詞nが出現する確率$P(t,n)$に基づくエントロピー$e(n)$は式(3)で求める．

$$e(n) = -\sum_t P(t,n)\log_2 P(t,n) \quad (3)$$

$e(n)$は任意の企業の統合報告書集合を 1 つの文書とみなし，それらの文書集合において名詞nが満遍なく出現している場合に高い値をとる尺度である．つまり$e(n)$に閾値を設けることにより特定の企業の統合報告書に出現する単語や，統合報告書のテキストデータ化の際に発生したノイズを除去することができる．

Step2 で得た$npmi$，及び Step3 で得たeが共に上位である名詞集合は E,S,G のいずれかに高い関連性があり，かつ会社名や人名などの ESG とは関係ない名詞を除いたものとなっている．

最終的に，表 2 に示すような特徴語を E,S,G それぞれに対して獲得できる．

表 2. 獲得する特徴語の例

トピック	特徴語
E	環境負荷，カーボン， 温室効果ガス，生物多様性
S	障がい者，育児休業， 地方創生，教育支援，男女
G	リスク管理，コンプライアンス， 透明性，法令遵守，是正措置

3.4. 学習データの自動生成

ページ単位で ESG の関連性の推定を行うためには，学習データをページ単位で用意しなければならないが，1 つのページにはタイトルや本文のような文字列が多く含まれるため，学習に必要な十分な量の学習データを手作業で作成するには多くの時間と労力が必要となる．そこで本研究では，統合報告書において ESG の特徴語が多く出現するページは，その特徴語のトピックに関連しているという考えに基づき，単語の出現情報を用いたルールを設定し，学習データを自動生成する．具体的には，統合報告書の学習用のページ集合において，E,S,G の各トピックについて，特徴語が重複無しで 5 個以上出現するページに対してそのトピックにラベル 1 を，特徴語が 1 つも出現しなかったページに対してそのトピックについてラベル 0 を付与する．ただし，学習データの精度を高めるために，上記のルールに少しでも当てはらないページは学習データから除く．例えば，あるページに G の特徴語が 1〜4 個出現した場合，そのページが G に関連するかの判定が困難であるため，G のラベルを付与せず，このページは学習データには追加しない．

ラベル 1 を付与するための条件である特徴語の閾値は，閾値を小さくすると学習データとして抽出できるページ数が増える一方でラベル付与の精度が下がり，閾値を大きくすると学習データとして抽出できるページ数が減る一方でラベル付与の精度が上がるというトレードオフの関係を考慮した結果，5 という値を選択した．

ルールベースにより抽出した学習データは，最終的に 3 次元の multi-hot 形式のラベルが付与される．ここで，ラベルの 1 次元目は E に関連するかどうかの 2 値(関連があれば 1，非関連であれば 0)を取る．同様に 2 次元目は S，3 次元目は G に対応する．

3.5. 分類モデル

本研究で扱うタスクにおいて，どのような分類器が適切であるかを比較するために，3 つのモデルを提案する．

3.5.1 モデル 1

入力を統合報告書のあるページとし，そのページが E,S,G のそれぞれについて関連するかを判別する分類器として，サポートベクトルマシン(SVM)を選択した．すなわち，統合報告書のページが E,S,G のそれぞれに関連するかそうでないかを判別する分類器を，E,S,G の 3 種類生成し，テストデータとなる統合報告書のページが E,S,G のぞれぞれに関連するか判定する(One vs Rest)．したがって，例えばページが E に関連するかを判別するための学習データは，E に関連するページが正例，E に非関連なページが負例（具体的には multi-hot ベクトルの 1 次元目が 1 のページが正例，0 のページが負例）となる．

各分類器の入力は[7]の素性選択手法を参考に，学習データから取得した素性および素性値を使用したベクトル(Bag-of-Words)とする．

3.5.2 モデル 2

入力を統合報告書のあるページとし，そのページが E,S,G のそれぞれについて関連するかを判別する深層学習によるマルチラベル分類器として多層パーセプトロン(MLP)を選択した．入力はモデル 1 で使用した 3 種類の分類器についての素性をまとめたものを素性としたベクトル(Bag-of-Words)とする．

モデルの入力層のノード数を入力ベクトルの次元数と同じとし，隠れ層は 1,000 次元が 3 層，300 次元が 3 層，25 次元が 3 層の計 9 層とする．出力層は E,S,G に対応する 3 次元とする．活性化関数はランプ関数(ReLU)を使用し，出力層はシグモイド関数を使用する．損失関数には BCE(Binary Cross Entropy)を適用する．

3.5.3 モデル 3

モデル 1,2 のようなベクトルを入力で受け取る分類器は，入力であるページに素性が含まれない場合，分類器の誤判定につながる可能性がある．そこで，入力を統合報告書のあるページとし，そのページに含まれる複数の文の文脈の情報を考慮して，かつ分類の手がかりとなる文を取捨選択し，そのページが E,S,G のそれぞれについて関連するかを判別するマルチラベル分類器として，双方向 LSTM(BiLSTM)と Attention を組み合わせた分類器を選択した．分類器の処理の流れは，統合報告書のページに含まれる各文を入力として受け取り，各文を単語埋め込み層により系列に変換し，双方向 LSTM に通すことで文の文脈を考慮した表現に変換，さらに Attention による重み付き和を計算することにより，ページの各文の情報を圧縮した表現を得る．最終的に圧縮した表現を線形変換することで E,S,G それぞれの関連性を推定する．単語埋め込み層の出力は 300 次元とし，双方向 LSTM の出力は 600 次元とする．出力層は E,S,G に対応する 3 次元とし，活性化関数はシグモイド関数を使用する．損失関数には BCE(Binary Cross Entropy)を適用する．

本研究では，あるページの入力に対して，ページに含まれる各文に対応する双方向 LSTM の出力を $H = (h_1, h_2, \dots, h_t)$としたとき，式(6)によりページに含まれる各文の情報をまとめた表現xを得る．

$$u_i = \tanh(W h_i) \tag{4}$$

$$\alpha_i = \frac{\exp(u_i^T v)}{\sum_i^t \exp(u_i^T v)} \tag{5}$$

$$x = \sum_i \alpha_i h_i \tag{6}$$

ただし，Wとvは学習により求められるパラメータであり，αは入力された文の注目度の重みを表す．

4. 実装と評価

本手法の評価を行うため，本手法を実装した．実装にあたり 2015 年から 2019 年までに発行された上場企業 412 社の 1,251 個の統合報告書 PDF ファイルの約 80,000 ページを学習データの自動生成および分類器の学習に使用した．

特徴語の抽出手法を実行することにより，最終的に E の特徴語を 187 個，S の特徴語を 176 個，E の特徴語を 152 個それぞれ抽出し，これらの特徴語を使用し，学習データを自動生成した，ただし E,S,G のどのトピックにも関連しないラベル[0,0,0]のページが他のラベルに比べて多く抽出されたため，分類モデルの学習の際にはデータの偏りを軽減するために，ラベル[0,0,0]のデータからランダムにサンプリングした 1,000 件をラベル[0,0,0]の学習データとする．

また，本手法で自動生成した学習データを用いて学習された分類器の性能を評価するための評価データとして，学習データの自動生成に使用していない 2020 年もしくは 2021 年に発行された 4 企業の統合報告書 PDF ファイルの計 372 ページに対して，各企業の web ページにて公開されている GRI スタンダード対照表を参考に，人手にて正解ラベルを付与し作成した．本手法で自動生成した学習データと，人手にて作成した評価データの，ラベルごとのページ

表 4. 各モデルの評価結果

	E(P)	E(R)	E(F)	S(P)	S(R)	S(F)	G(P)	G(R)	G(F)
モデル 1	87.6	77.3	82.1	71.5	72.7	72.1	68.3	71.7	69.9
モデル 2	78.7	80.7	79.7	62.6	75.8	68.6	69.9	77.4	73.4
モデル 3	78.4	88.2	83.0	64.0	82.0	71.9	65.8	82.4	73.2

表 5. 学習データの評価結果

	E(P)	E(R)	E(F)	S(P)	S(R)	S(F)	G(P)	G(R)	G(F)
ルール	92.5	97.4	94.9	92.5	92.5	92.5	92.5	100	96.1

数を表 3 に示す．ただし，〇は該当するトピックに関連することを示す.

表 3. 自動生成した学習データと，人手にて作成した評価データのページ数

ラベル	E	S	G	学習データ	評価データ
[0, 0, 0]				13870	89
[1, 0, 0]	〇			375	62
[0, 1, 0]		〇		797	57
[0, 0, 1]			〇	923	74
[1, 1, 0]	〇	〇		97	5
[1, 0, 1]	〇		〇	203	22
[0, 1, 1]		〇	〇	1228	36
[1, 1, 1]	〇	〇	〇	628	27

提案モデル 3 は，入力としてあるページに含まれる複数の文を受け取るが，ページごとに含まれる文の数は異なるため，モデルに入力する文の数を統一する必要がある．本手法では入力文の数の上限を 100 とした．ただしページに含まれる文については，分類の判断に悪影響を与える可能性のある明らかなノイズを予め除去するため，以下の条件を全て満たす文字列とする.

・文字数が 4 以上かつ 220 以下
・“http”を含まない
・同じ文字の連結ではない（¥¥¥¥¥など）
・同じページに重複して出現していない

本手法の評価を行うために，評価データを使用し，3 つの提案モデルについて，モデルの出力である 3 次元ベクトルの次元ごとに 2 値分類で評価を行った．評価指標として適合率(P)，再現率(R)，F1-score(F)を求めた．表 4 に各提案モデルに対する評価値を示す．表 4 の見方については，先頭に“E”のついた評価指標は出力である 3 次元ベクトルの，1 次元目における 2 値分類の評価を表しており，同様に“S”は 2 次元目，“G”は 3 次元目における 2 値分類の評価を表す.

5. 考察

以下の 3 つの内容に対して考察を行う.

(1)学習データを自動生成したことによる分類性能への影響
(2)形式崩れの多いテキストデータに対して本研究のようなタスクを行う場合，どのような機械学習モデルが適切なのか
(3)崩れたテキストデータに対しての機械学習モデルによる文単位での処理の有効性

まず(1)について述べる．本研究では，ルールベースにより学習データを自動生成しているが，モデルの適切な学習のためには，学習データのラベル付与の精度が重要であり，モデルの分類性能に影響する．そこで，本手法で自動生成された学習データの評価を行った．自動生成された学習データは，18,121 ページと数が膨大であるため，ルールにより付与されたラベルごとにランダムに 10 件選んだ，合計 80 ページに対し，人手で確認し正解ラベルを付与することで，学習データの定量的な評価を行う．ルールにより自動生成したラベルと人手により付与した正解ラベルを用いて，ESG の各トピックごとの 2 値分類の適合率(P)，再現率(R)，F1-score(F)を表 5 に示す.

表 5 より，各トピックごとに，ルールによる学習データのラベル付与の精度を確認すると，高い精度でラベルを付与できているが，表 4 の評価データを用いた提案モデルによる評価値は，学習データの評価値と大きな差がある．これは，ルールに当てはまるページに対しては正しくラベル付与し学習データとして用意できたが，ルールに当てはまらなかったページに対してモデルは学習できず，適切な分類が出来なかったためと考える.

また，本研究の学習データの自動生成手法におい

て学習データのラベル付与の誤りや見逃しが起こる原因としては，ラベル付与に単語の出現情報を利用していることが起因していると考える．まず，ルールにより誤って関連ありと判断される場合(0→1)について考察する．本手法では学習データを自動生成する際にページの内容については考慮せずに，あるページに特徴語が一定数出現したかどうかでそのトピックに関連があるか判断するため，ESG に非関連なページに対しても，特徴語を含んでいると ESG に関連があると誤認識してしまう場合がある．この問題が頻繁に発生するのが目次ページである．目次ページは，ESG 情報や他の内容が PDF のどの位置に掲載されているか示すページであり，目次ページ自体に ESG 情報は含まれないため ESG に非関連と認識することが望ましいが，本手法のルールでは ESG に関連ありと誤認識されてしまう．このようなページが学習データに追加されると，分類モデルの適合率の低下に繋がる可能性があるため，学習データから極力取り除くべきである．

次にあるルールによるラベル付与の見逃し(1→0)について考察する．我々は特徴語を一定数含むページを学習データとしたが，本手法で全ての特徴語は獲得できていないため，獲得できなかった特徴語に関する話題のページについては，そのトピックのラベルを付与できない．例えば，環境に配慮した取り組みを表す“グリーン調達”という単語は E の特徴語として適切だが，本手法では特徴語として抽出できなかったため，あるページに“グリーン調達”が出現しても E に非関連と認識されてしまう場合がある．このようなラベル付与を見逃したページは分類モデルの再現率の低下を招くと考えられるため，学習データに追加させないほうがよい．以上より，分類モデルの性能を向上させるには，本手法でのルールベースによる学習データの自動生成手法の改善が必要であることが分かった．

次に(2)について述べる．従来，本タスクで扱うような形式が崩れたテキストデータに対して機械学習を適用する場合は，単語単位に分割して行われてきた．これは，文が崩れても単語の出現情報は失われにくいという点で有効だが，単語の順序情報を失うという欠点も存在する．本研究では，入力に単語の出現情報を利用した分類器として提案モデル 1,2 を，文の文脈情報を利用した分類器として提案モデル 3 を選択し，形式崩れの多いテキストデータに対してどのような機械学習モデルが適切かの考察を行う．

まず，各提案モデルによる統合報告書のページへのラベル付与の正確性を評価するため，表 5 の適合率に注目する．表 5 より適合率は，全ての提案モデルで E が 8 割，S と G は 6,7 割となっており．モデルによる適合率の大きな違いは見られなかったが，ESG の全てのトピックで最も高い適合率を得たのは提案モデル 1 であった．一方で，提案モデル 1,2 に比べて，提案モデル 3 は適合率が若干低い傾向にある．これは，モデルの特徴に起因すると考える．提案モデル 3 は入力であるページの各文を受け取るため，提案モデル 1, 2 よりも受け取る情報が多くなる．統合報告書の ESG 情報を含むページには，ESG と関連がない話題が含まれていることもあるため，広範囲の情報を受け取る提案モデル 3 は，学習時にそのような話題についても ESG の特徴と学習したため，他のモデルに比べ適合率が下がったと考える．

次に各提案モデルが統合報告書のページに対して，ラベル付与をどの程度取りこぼしていないかを考察するため，表 5 の再現率に注目する．表 5 より，全ての提案モデルが ESG の全てのトピックにおいて 7 割を超えており，ESG 情報を比較的見落とすこと無く認識できている．特に，提案モデル 3 は ESG の各トピックについて再現率が 8 割を超えており，他の提案モデルよりも多くの ESG 情報を認識している．提案モデル 1,2 は入力としてベクトルを受け取るため，入力ページをベクトルに変換する必要があるが，変換する際にページに含まれる文の文脈情報は無視され，かつ学習に使用した素性以外の単語の出現情報は無視される．そのため細かな ESG 情報を認識しにくい一方で，提案モデル 3 は入力として，入力ページに含まれる各文を受け取るため，各文の文脈情報を捉えることができ，ESG 情報の取りこぼしが減り再現率が高い結果になったと考える．ただし，単語の出現情報を利用するモデルは情報の見落としにより再現率が低くなると予想していたが，そのような提案モデル 1, 2 において，全てのトピックで再現率が 7,8 割を達成しているため，使用した素性が有効であったと考える．

以上の考察より，形式崩れの多いテキストデータを用いる本タスクにおいて，ESG の関連性推定の正確性を優先する場合は，提案モデル 1, 2 のような単語の出現情報を用いたモデルが適切であり，ESG 情報の取りこぼしを減らすことを優先する場合は，提案モデル 3 のような文の文脈情報を用いたモデルが適切な選択であると考える．

最後に(3)について述べる．本研究では崩れたテキストデータから ESG 情報をページ単位で抽出する際に，文単位での処理が有効であるかを調査するために提案モデル 3 を選択した．提案モデル 3 は，入力情報に対して分類の手がかりとしての重みを計算することで，情報の取捨選択が可能な Attention を持つ．ノイズが多いデータを文単位で扱うためには，不要な文をフィルタリングし，重要な文を強調して

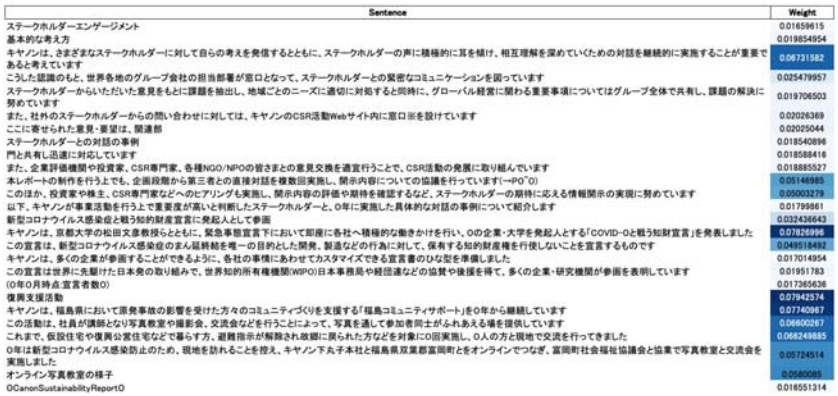

図 2. 統合報告書のページを提案モデル 3 に入力したときの Attention による文の注目度の例.
正解ラベル[0,1,1], 予測ラベル[0,1,1]

処理を行う必要があると考え Attention を採用したが，本タスクにおいて Attention が有効なのかを調査した．図 2 に，統合報告書のページが入力されたときの，Attention による文の重みを示す．図 2 は，復興支援活動(S)とステークホルダーエンゲージメント(G)に関連するページを入力した際の文の注目度である．「復興支援活動」や「キャノンは、さまざまなステークホルダーに対して...」のような重要な文と,「門と協力し迅速に対応します」や「基本的な考え方」のような分類の手がかりにならないような文を正確に判別できており，このページに対して G,S に関連があると正しく推定できている．この例は，Attention を用いることで，途切れた文が含まれるようなノイズの多いテキストデータでもある程度正確に文の重要性を判別することができることを示しており，Attention により提案モデル 3 は高い再現率を得たと考える．また，提案モデル 3 のようなアーキテクチャは統合報告書だけでなく，他の崩れたテキストデータを対象にした文書やページ単位の分類タスクにおいても有効と考える．

6. まとめ

本研究では，企業の ESG 情報の開示媒体の 1 つである統合報告書から，ESG 関連ページを自動で推定する手法を提案した．本タスクでは扱うデータは形式の崩れたテキストデータであることを考慮し，複数の機械学習モデルを提案し，適切なモデルについて考察を行った．その結果，ESG の関連性推定の正確性を優先する場合は単語の出現情報を利用したモデル，ESG 情報の網羅性を優先する場合は，文の文脈情報を利用したモデルが適切であることを示した．

今後は，学習データの自動生成手法を見直し，精度を改善することで分類モデルの性能向上を目指す．

参考文献

[1] Global Sustainable Investment Alliance: GLOBAL SUSTAINABLE INVESTMENT REVIEW 2021. http://www.gsi-alliance.org/wp-content/uploads/2021/08/GSIR-20201.pdf

[2] 年金積立金管理運用独立行政法人（GPIF）: 2020 年度 ESG 活動報告. https://www.gpif.go.jp/investment/GPIF_ESGReport_FY2020_J.pdf

[3] 一般社団法人生命保険協会: 企業価値向上に向けた取り組みに関するアンケート集計結果一覧（2020 年度版）企業様向けアンケート. https://www.seiho.or.jp/info/news/2021/pdf/20210416_4-3.pdf

[4] 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, 第 26 回 金融情報学研究会, 2021.

[5] 中尾悠利子, 石野亜耶, 岡田斎: ニューラルネットワークによるサステナビリティ情報のテキスト分析――経営トップメッセージの環境・社会記述分析への適用, 企業と社会フォーラム学会誌, 第 8 号, pp.57-72, 2019.

[6] 高野海斗, 酒井浩之, 中川慧: 学習データの自動生成による深層学習を用いた株主招集通知の重要ページ抽出, 人工知能学会論文誌 2021 年 36 巻 1 号 p.WI2-G_1-19.

[7] 酒井浩之, 松下和暉, 北島良三: 学習データの自動生成による決算短信からの業績要因文の抽出, 日本知能情報ファジィ学会誌, Vol.31, No.2, pp.653 - 661 (2019).

[8] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, Eduard Hovy: Hierarchical Attention Networks for Document Classification, Proceedings of NAACL-HLT 2016, pages 1480–1489.