投資支援のためのニュース記事からのESG関連文抽出

ESG-related text extraction from news articles for investment support

吉田　朋弘[1]　小澤　誠一[1]*
Tomohiro Yoshida[1]　Seiichi Ozawa[1]
渡辺　一男[2]　廣瀬　勇秀[2]　池田　佳弘2
Kazuo Watanabe[2]　Takehide Hirose[2]　Yoshihiro Ikeda2
飯塚　正昭[2]　西田　大輔[2]
Masaaki Iizuka[2]　Daisuke Nishida[2]

[1] 神戸大学大学院工学研究科
[1] Graduate School of Engineering, Kobe University
[2] 三井住友DSアセットマネジメント株式会社
[2] Sumitomo Mitsui DS Asset Management Company, Limited

Abstract: 近年ESGという言葉が急速に普及し、企業は企業の長期的な成長に影響するESG活動という環境や社会に配慮した活動を求められるようになった。そこで従来の投資手法である売上高や利益などの財務指標のみを重要視した投資手法ではなく、ESGへの取り組みという非財務情報の要素も考慮した経営を行う企業に投資する、「ESG投資」が台頭してきている。本研究では、実際に企業が行っているESG活動の動向や実績などの記事をロイターニュース記事から取得し、どの企業に投資するのがよいのかという判断材料を作ることが目的となる。具体的には、ロイター記事に対してアノテーションを行いロイター記事データセットを作成し、作成したデータセットに対してBERTのファインチューニングを行い、ESGの文章分類を行うことでESG関連文の抽出を行った。その結果、BERTのファインチューニングモデルは高い性能を発揮した。またSHAP値による判断根拠となる単語を可視化したことでモデルの有効性を示せたとともに、重要語の抽出が行えた。

1　はじめに

環境 (Environment), 社会 (Social), コーポレートガバナンス (Governance) という3つの非財務情報的観点を投資判断に取り入れるESG投資は長い目で見ると合理的であるという考えが広まりつつある。日本でESGへの関心が増え始めたのが2015年からといわれており、この理由として2015年は地球温暖化防止を防止するための国際的な枠組みである「パリ協定」が締結された年であり、日本では2015年に年金積立金管理運用独立行政法人（GPIF）が国連の責任投資原則（PRI）に署名した年であるからと考えられる。GPIFは日本の公的年金運用機関であり約200兆円という巨額な資産を運用しており [1]、金融市場では大きな影響力を持っている。 PRIに署名した団体は投資分析と意思決定プロセスにESGの視点を組みこむことや、投資対象の主体に対してESGの主体に対してESGの課題について適切な開示を求められるので、ESG投資に真摯に向き合わなければいけなくなる。

ESG投資に向き合うためには、ESG投資の投資先となる企業のESG活動を知らなければなりません。企業は統合報告書、CSRレポートやサステイナブルレポートを通じて自社のESG活動における情報を包括的に開示する。しかし、これらのレポートを作成、開示するには時間がかかりタイムリーな情報の取得が困難である。そこで、ニュース記事からのタイムリーにESG情報の収集を目指す。

そのため、本稿では金融情報の提供に強いロイター・ニュース記事を用いてESG情報記事の抽出をおこなう文章分類器を構築することを目標としている。ニュース記事は日々様々なトピックが付加される流動性が高い情報である。そのような中からESG関連記事を抽出する試みの第一段階として、特定の期間におけるロイター・ニュース記事を用いて文章分類器の作成を行った。

*連絡先：神戸大学大学院工学研究科
〒 657-8501 神戸市灘区六甲台町 1-1
E-mail: ozawasei@kobe-u.ac.jp

2 関連研究

ESG に関連文章抽出に関する研究は企業が自社の ESG 活動を記述している有価証券報告書からの抽出 [2] や統合報告書からの抽出 [3] が存在する。[2] 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などの有価証券報告書において ESG 関連文章が存在するであろう項目文に対して GRI スタンダード [4] に基づいたアノテーションを行い BERT を用いたテキスト分類を通じて、ESG 関連文を抽出する手法を提案した。また、[3] は統合報告書から ESG 関連文の抽出を目指した。E,S,G が全体としてどのような話題が該当するか投資家によって変わり、明確に定められていないとして、ESG 文章データセットを ESG に高い関連性がある特徴語を設定し、特徴語の出現頻度に従った学習データの自動生成をした。生成したデータセットに対して、E,S,G の 3 値のマルチラベル分類を行っている。

これらの関連研究のように企業の公的文書から ESG 関連の文章を抽出しようと試みるタスクは研究が進んでいる。一方、ニュースデータに着目した ESG 関連の応用例は少ない。ニュースデータを用いた ESG 関連研究として、Word2Vec を用いてニューステキストから ESG トピックワードを抽出し、企業の ESG 評価を行うことでその評価に基づいた ESG ファクターを分析した [5] がある。本研究では [2] に準じてニュース記事の ESG 分類を行った。

3 手法

3.1 使用データ

日本語ロイター・ニュース記事の 2021 年度 1 月 3 日から 2021 年度 5 月 13 日までに提供されている、ロイター・ニュース記事本文のテキストデータ用いた。ロイター・ニュースとは、トムソン・ロイター通信社が配信しているリアルタイムな速報ニュースであり、国内の株式関連記事や、市況、為替、金利などの金融関係についての情報が配信されている。

3.2 前処理

収集したロイター・ニュース記事の中には ESG に関する記事や「今週の焦点」や「アングル」など今週で注目に値する経済動向をまとめた記事や数字情報が多い記事である「連結決済記事」や株式の「新規上場日程」が一覧になっている記事などの様々な記事が混在している。

他記事と比べて数字情報が多くを占める記事では前処理で数字を「0」に置換したとき、記事に多くの「0」が出現する記事となってしまい ESG の分類上不要なノイズ記事になってしまう可能性が高い。そこで ESG に関係のない、または複数のトピックが入ってしまい分類するのが困難な記事などを Headline 情報に基づいて削除することで精度向上を図った。

Headline 情報というのは記事の見出しにあたる情報であり、特定の記事をルールベースで取得できる。Headline 情報に従って扱うテキストデータを選択した後、GRI スタンダード [4] に基づいたアノテーションを行うことでロイター・ニュース記事の ESG データセットの作成を行った。以下に Headline 情報に基づいて削除した項目を示す。

- **数字情報が多い記事**

「新規上場日程」、「決算市場予測」、「配当予想」など

- **コラムやまとめの記事**

「COLUMN」、「アングル」、「週の焦点」など

以上の単語が Headline に含まれている記事をデータセットから取り除いた。

また、文章に対して行った前処理として

- 数字の 0 置換
- アルファベットの小文字化
- ( ),[ ],<>で囲まれた文章

を行った。上記括弧で囲まれた文章は参考 URL や細かい補足説明などが記述されている。

3.3 GRI スタンダード

アノテーションに用いた GRI スタンダードは、GRI(Global Reporting Intiative) というサステナビリティに関する国際基準策定を行う非営利団体が策定した、経済、環境、社会に与えるインパクトを一般に報告する際の、グローバルレベルにおけるベストプラクティスを提示するための規準である。GRI スタンダードに基づいて作成されたサステナビリティ報告書では、組織が持続可能な発展に対して与える情報が提供される。

3.4 モデルの構築

E,S,G, その他の 4 クラス分類を行うために、どのような文章埋め込みが適切であるのかを比較するために、文章の埋め込み表現モデルとして TF-IDF,Word2Vec,BERT の 3 種類を比較した。

- **TF-IDF**

TF-IDF[6] は Bag-of-words の手法であり、1 つの文章中の単語出現頻度に全文章中における単語の出現頻度を考慮したものである。TF-IDF を比較手法として用いた理由としては、ESG 関連文には「気候変動」や「雇用」などの ESG を表す特徴的に出現する単語が存在しているため、有効な手法であると考えたためである。ま

た TF-IDF による文章ベクトル化はスパースなベクトルとなることが多いため、主成分分析 (PCA) による次元削減を行い累積寄与率が 80%となる次元の特徴量を用いた分類も行った。TF-IDF[6] の実装では、Mecab と NEologd 辞書を用いた単語分割を行い、作成したコーパスを持ちて Word2Vec[7] を学習させたのち、サポートベクトルマシンを用いた分類を行った。

・ **Word2Vec**

Word2Vec[7] は単語の共起関係に着目し、コンテキストに依存しない埋め込みベクトルを作成する。各単語につき 1 つのベクトル表現を作成し、単語の埋め込み表現を獲得する。Word2Vec を用いたモデルでは、各記事の入力単語を Word2Vec を用いた埋め込み表現で表し、各単語の平均ベクトルを算出することで各記事の文章ベクトルを得ている。Word2Vec[7] の実装においては作成したロイター・ニュースデータセットを Mecab で分かち書きし、各記事のコーパスを作成し、作成したコーパスを持ちて Word2Vec[7] を学習させたのち、サポートベクトルマシンを用いた分類を行った。

・ **BERT**

BERT[8] は単語が使用されているコンテキストに基づいた単語埋め込み表現を獲得でき、異なるコンテキストにおいて同じ単語に対して複数のベクトル表現を持つことを可能にしている。BERT[8] の実装においては入力文章の頭につける [CLS] トークンに対応する transformer encoder の最終層のノードを用いて分類問題の予測をした。文章分類器には全層結合の 1 層の全結合層を用いた。ファインチューニング時にはロイター・ニュースデータセットを用いて、学習用データの損失が 7 回下がらなくなると Earlystopping をかけて学習をストップさせた。使用した重みは損失が最小となる Epoch の重みを用いている。BERT の実装においては日本語 BERT の学習済みモデルとして、東北大学の乾・鈴木研究室が公開しているモデル [9] を利用した。

4 実験評価と考察

いずれの分類手法が有効であるのかを確かめるために、評価実験を行った。評価手法に関しては、ESG 関連文とその他の文章を比較して ESG と関連しないその他の文章が多いという不均衡性を考慮して F 値を用いた。ニュース記事からの抽出タスクであるので、抽出できないという機会損失をなくしたいために recall 値も重要な値であると考えられるので recall 値も示しておく。

実験結果を表 1 と表 2 に示す。そのなかではファインチューニングありの BERT モデルが最もよい値を示した。次いで、TF-IDF モデルが良い性能を示した。各実験において BERT をファインチューニングしたモデルを用いた方が macroF 値において最も高くなった。比較実験の中で Word2Vec モデルの F 値が他のモデルより F 値が低くなった原因として、学習データ数が少なく、十分な学習が行われていなかったからであると考えられる。また、TF-IDF のモデルが比較的良い性能を示したのは、用いたロイター・ニュース記事の ESG 関連データセットの期間が約 6 か月であり、ニュース記事の特徴である、ある記事に対して関連した類似の記事が出版されるということが起きているため、同様のトピックの文章が多かったためと考えられる。

表 1: 各モデルの F 値の比較.

モデル	E	S	G	その他
TF-IDF	0.83	0.85	0.94	0.96
TF-IDF(PCA)	0.86	0.85	091	0.96
Word2Vec	0.0	0.32	0.0	0.86
BERT ファインチューニングなし	0.87	0.73	0.89	0.95
BERT ファインチューニングあり	0.92	0.91	0.96	0.98

表 2: 各モデルの Recall の比較..

モデル	E	S	G	その他
TF-IDF	0.71	0.74	0.94	0.99
TF-IDF(PCA)	0.75	0.74	0.91	0.99
Word2vec	0.0	0.19	0.0	1.0
BERT ファインチューニングなし	0.83	0.69	0.85	0.96
BERT ファインチューニングあり	0.96	0.86	0.98	0.98

4.1 SHAP 値による可視化

ブラックボックスになることが多い機械学習モデルにおいて、なぜその予想を行ったのかを表すために判断根拠の可視化というのは重要である。本稿では機械学習モデルの判断根拠の可視化のために SHAP 値 (SHapley Additive exPlanations)[10] を用いた。SHAP 値は協力ゲーム理論において Shapley 値を機械学習に応用し、モデルに入力される各説明変数がモデル予測に与える貢献度を評価しようとするものである。

ここでは、最も性能が良かった BERT ファインチューニングモデルの E,S,G, その他の 4 値分類においていずれかの単語が分類に影響を与えているのかを調べた。具体的にはテストデータにおいて分類予想が正しかったデータに対して、SHAP 値によるいずれかの単語が分類に影響を与えているのかを上位 20 単語抽出した。

図 1: 環境カテゴリへの分類寄与が高い単語.

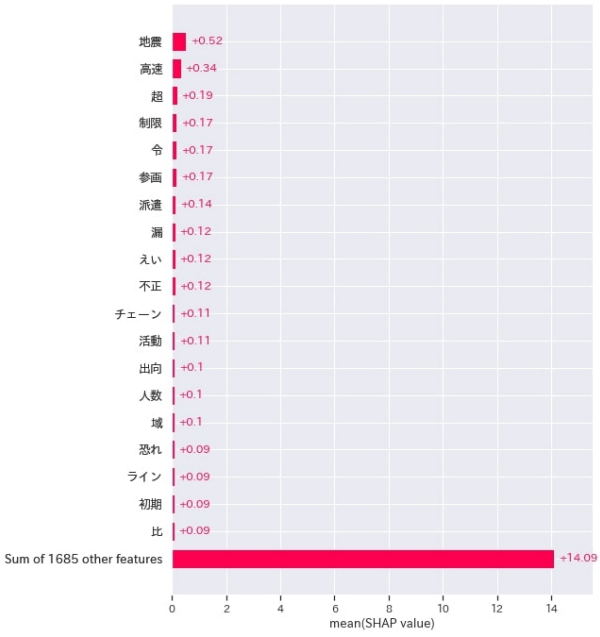

図 2: 社会カテゴリへの分類寄与が高い単語.

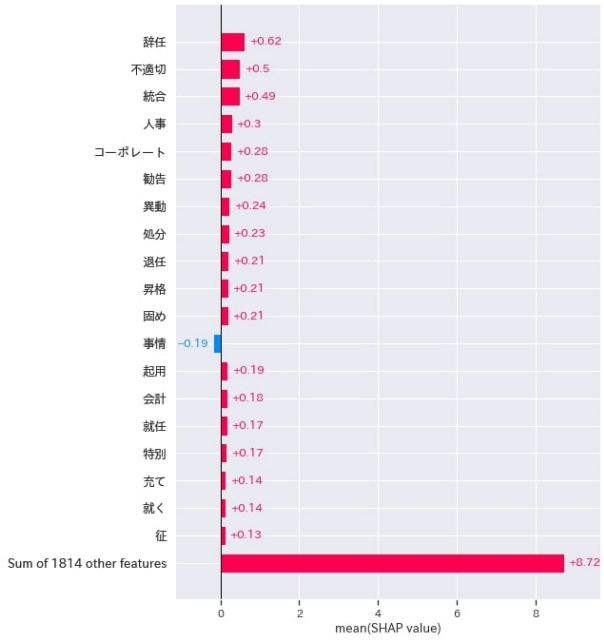

図 3: ガバナンス項目への分類寄与が高い単語.

抽出した影響度の高い上位 20 単語を図 1、図 2、図 3 にそれぞれ示す。これらの単語はそれぞれ環境 (E)、社会 (S)、ガバナンス (G) の分類に影響を持っている単語群である。抽出された単語を見てみると、環境カテゴリではクリーン発電関係と思われる「風力」や「洋上」といった単語や、脱炭素関係だと考えられる「燃料」、「炭素」、「ハイブリッド」等の単語が見られる。社会カテゴリでは「地震」がトップになっているがこれは意外な結果であり、2021 年 2 月の福島県沖地震でのサプライチェーン工場の停止の影響を強く受けていると考えられる。ガバナンスカテゴリでは、役員進退の「就任」など、コーポレートガバナンス関連の単語が見受けられる。

5 さいごに

本稿では、ニュース記事からの ESG 関連記事の抽出を目標として、ロイター・ニュース記事にアノテーションを行い TF-IDF, Word2Vec, BERT の各文章埋め込みモデルを用いて、E,S,G, その他の 4 値分類を行った性能を評価した。結果としては TF-IDF モデルと BERT のファインチューニングモデルは比較的高い F 値を示した。また、判断根拠の可視化とし SHAP 値による可視化を行った。判断根拠の可視化の部分においては、BERT のファインチューニングモデルにおいて、いずれかの単語が分類に大きな影響を与えているのかを上位 20 単語文を可視化した。ESG に関連があるような単語が抽出できており、また社会項目では時事関連の単

語の抽出もできていると見れる。モデルの有効性が確かめられたと考える。本稿が示めした、SHAP 値による可視化での判断根拠となる単語抽出や BERT,TF-DF モデルの有効性は ESG 関連文に含まれる、ESG の特徴後の抽出を通じたルールベースのラベル付与などに応用できると考えられる。今後の課題として、ニュース記事においては日々新たなトピックが出現するので、現在の学習データセットに入っていないデータに適応させる必要がある。新たに出現する未知のトピックに対応するべく現在の分類問題をオープンワールド問題に拡張することが大きな課題になる。また、本稿では COLUMN 等の記事は除外しているが、それらの記事の中にも ESG 関連文章が存在する可能性があるため、記事全文での分類ではなく、記事内の文章単位でも分類できるように拡張することを課題としている。

参考文献

[1] 年金積立金管理運用独立行政法人.: 2021 年度の運用状況, https://www.gpif.go.jp/operation/the-latest-results.html

[2] 土橋良太, 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

[3] 河村康平, 高野海斗, 酒井浩之, 永並健吾, 中川慧.: 機械学習を用いた統合報告書の ESG 関連ページ推定, 第 27 回 金融情報学研究会, 2021

[4] Global Reporting Initiative.: GRI サステナビリティ・レポーティング・スタンダード, https://www.globalreporting.org/how-to-use-the-gri-standards/gri-standards-japanese-translations/

[5] 秋山祥吾, 江口潤一, 鈴木智也 Word2Vec を用いたニューステキストの ESG ファクター運用 *The 34th Annual Conference of the Japanese Society for Artificial Intelligence*, 2020

[6] G. Salton, M. McGill.: Introduction to Modern Information Retrieval, *In McGraw-Hill*, 1983

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean.: Efficient estimation of word representations in vector space, *International Conference on Learning Representations*, pp. 1–12, 2013

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova.: BERT: Pre-training of deep bidirectional transformers for language understanding, *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long and Short Papers), pp. 4171–4186, 2019

[9] 東北大学 乾・鈴木研究室.: Pretrained Japanese BERT models, https://github.com/cl-tohoku/bert-japanese

[10] Scott M. Lundberg, Su-ln Lee.: A unified approach to interpreting model predictions, *In NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768–4777, 2017