ESG スコアの判断の根拠となるテキストの抽出

河南 直希

近年，投資，企業経営の分野で「ESG」という概念への関心が急激に高まっている．企業は環境保護のような短期で利益をもたらさない活動を求められ，それらを行っている企業への優先的な投資が盛んに行われるようになっている．企業の ESG 情報は評価会社によりスコア化されているが，既存スコアの多くは根拠が不明瞭であり，細かい算出手法はノウハウ等で多くが公開されていない．また既存の調査によると，企業の理解よりも多くの投資家が ESG 情報の開示が不十分と考えている．そこで本研究では，ESG 関連文の属性を複数定義し，それらを組み合わせて有価証券報告書のテキストから企業の ESG 情報を自動で抽出することを目的とする．

本研究では，深層学習モデルで有価証券報告書の文を分類するが，企業の ESG 情報の判断には従来の ESG 関連のテキスト分類の研究のような特定の観点による分析のみでは不十分である．ESG 分類の他にも極性分類など複数の属性を文に自動で付与することで，ESG スコアの根拠として考えられるテキストを抽出できる．属性が「環境（E)」かつ「肯定」のテキストの出力は，高い環境（E）スコアの根拠となり得，同様に属性が「環境（E)」かつ「否定」であるテキストは低い環境（E）スコアの根拠となり得る．このように本研究では深層学習モデルにより自動分類されたテキストに基づき，既存の ESG スコアの根拠となっているか否かを判断する．

本研究では，BERT をファインチューニングすることで各属性の分類モデルを構築する．さらに，データ数の少ないラベルにのみ追加でアノテーションしデータ拡張を行うことで，ESG 情報に対するデータ拡張の有効性を検証する．有価証券報告書からの ESG スコアの判断の根拠となるテキストの抽出のため，初めに人手でアノテーション作業を行い，ESG 関連文データセットを作成する．テキストは，金融庁の開示システム EDINET から収集した XBML データを解析・整形し，さらにタイトルや小見出し，記号などの不要な文字を削除してデータセット作成した．本データセットで扱う文数は，2 業種 15 社の計 4,532 文である．作成した ESG 関連文データセットは，全ての属性において Fleiss の κ 係数を用いたアノテーション作業者間の一致度が 0.6 を超えたことから，判定者に依存しない妥当な基準で作成されたことが示された．

提案手法を用いた実験では，ESG 関連文の分類モデルを構築する際，データ数の少ないラベルに絞ってアノテーションし，データ拡張することで，人的資本ラベルの 5 分割交差検証を用いた F 値（F1-score）が 0.780 となり，他のラベルも全て F 値が 0.8 以上と精度が改善されたため，ESG 関連文データセット作成における少数ラベルのデータ拡張の有効性を示した．また，提案手法の分類モデルを用いて未知の企業について属性ラベルを分類し，ESG スコアの判断の根拠となるテキストが抽出可能であることを示した．文意に企業統治を含む文は分類精度 0.8 以上で正しく分類され，ESG スコアの判断の根拠となるテキストが抽出できた．一方で一部の文は環境と社会資本の両方の属性の根拠が含まれ，誤分類となってしまいスコアの根拠としては不適切な文が抽出された．よって複数の属性を含むテキストの扱いが今後の課題である．

（指導教員　関 洋平）