# 経済記事からの不祥事報道検知

Jason Bennett* 野原 崇史 * Fei Cheng+ 石田 隆 * 宮尾 祐介 +
* 三井住友アセットマネジメント + 国立情報学研究所 金融スマートデータ研究センター
*{jason_bennett, takafumi_nohara, takashi_ishida}@smam-jp.com, +{fei-cheng, yusuke}@nii.ac.jp

## 1 はじめに

**概要** 本論文では、経済記事を二値分類するための実務的な End-to-End のシステムの設計・構築を行なった。分類システムを実運用に向けてデザインする上では精度は重要な指標の一つに過ぎず、解釈性、頑健性等に関する課題および解決策に関する研究を行なった。時間の経過とともに正のクラスの定義が変化する実環境の難しさにも対応するため、再学習の仕組みも取り入れた。

**実務的背景** 資産運用実務において近年、非財務情報として環境（Environment）、社会（Social）、企業統治（Governance）への取り組み姿勢を評価に取り入れ投資する ESG 投資に注目が集まっている。[1]

ESG 投資において、企業における「情報漏洩」「リコール」「セクハラ」などのネガティブイベントを速やかに把握することは重要である。2017 年だけを見ても大型銘柄の「偽装」「隠蔽」「改ざん」などのニュースが相次ぎ報道され、ファンドマネジャーがこれらの不祥事イベントを早期に認識し、企業不祥事イベントに関する株価・財務内容への影響分析や投資分析を行うことの重要性が増加している。

**システム構築の主な課題** 多岐に渡る不祥事イベントを正しく早期に認識するためには、ESG 投資の担当者が日々大量の経済記事に目を通し、その中から不祥事関連の記事を特定する必要がある。しかし、これには莫大な時間がかかる上に、不祥事関連記事の割合は1%未満と僅かであり、ほとんどは不祥事の情報収集に役立たない。珍しいだけではなく、不祥事記事の見逃しも発生しやすいため、ファンドマネジャーが不祥事の分析に出遅れることも課題であった。

そこで我々は経済記事を「不祥事」「不祥事ではない」に分類する、スケーラブルで End-to-End な分類器システムを設計･構築した。経済記事の分類に関する研究はセンチメント分類やイベント検出 [3] など様々行われてきたものの、不祥事検知についての研究は我々の知る限り研究されていない。我々の研究により ESG 投資の担当者やファンドマネジャーが、「不祥事」と分類予測された記事のみを閲覧し、日々の情報収集を短時間で済ませることに道を開いた。

分類器の学習においては適合率や再現率などの精度が注目されがちだが、実務上はその他にも考慮しなければならない様々な課題が浮上し、その中でも重要な 3 点を取り上げる。

1 つ目は解釈性の担保である。深層学習など複雑なモデルを利用することは分類精度への多少の貢献があるかもしれないが、分類システムの出力結果をユーザであるファンドマネージャーや顧客である年金基金等に直感的に、もしくは簡易に説明することは非常に難しく、精度と解釈性のトレードオフが存在する [1]。そのためモデルにはロジスティック回帰、記事の特徴量としては $N$-gram を利用した。結果的にシステム担当者自身が分類器で重要視されている特徴量を調べやすくなり、品質維持の上でも好ましいシステムとなった。

2 つ目は不祥事に関するデータの分布の変化がある状況下で、システムにある程度の頑健性を持たせることである。不祥事分類器を学習する上で、特定の期間のみに対して有効な特徴が頻繁に観測されたが、これらを含めることは予測期間に対する汎化性能の劣化に繋がることがわかった。そこで「固有名詞」を除くといった特徴調整が効果的であることがわかった。

3 つ目は再学習の必要性である。時間の経過とともに、絶えず新たな種類の不祥事が登場する環境でも、システムのパフォーマンスを維持したい。そこで、ある記事に対するシステムの分類結果が誤っているとユーザが判断した場合は、それをユーザインターフェース（UI）から変更し、分類器の学習に再度用いる仕組みを構築した。本論文では再学習の擬似的なシミュレーションを行い、その効果も検証した。

次章以降の本論文の流れは次の通りである。第 2 章では用いたデータやその前処理、すべての実験に共通

[1] 例えば、世界最大の年金基金である年金積立金管理運用独立行政法人（GPIF）が 2017 年に 1 兆円規模で ESG 投資を開始した。

表 1: 利用したモデルの一覧。$N$-gram は $N = 1$（Uni-gram）と $N = 2$(Bi-gram) の 2 種類、記事の長さは全文と冒頭 2 文のみの 2 種類を用いた。特徴調整を行う場合は記号、数、助動詞、一部の助詞、固有表現などを除いた。

| モデル # | $N$-gram | 記事長さ | 特徴調整 |
|---|---|---|---|
| モデル 1 | $N = 1$ | 全文 | 無し |
| モデル 2 | $N = 1$ | 冒頭 2 文 | 無し |
| モデル 3 | $N = 1$ | 全文 | 有り |
| モデル 4 | $N = 1$ | 冒頭 2 文 | 有り |
| モデル 5 | $N = 2$ | 全文 | 無し |
| モデル 6 | $N = 2$ | 冒頭 2 文 | 無し |
| モデル 7 | $N = 2$ | 全文 | 有り |
| モデル 8 | $N = 2$ | 冒頭 2 文 | 有り |

する設定等について説明し、第 3 章〜第 5 章では解釈性・頑健性のある不祥事分類システムの設計・構築、および分類器の再学習に関する実験とその結果を示す。最後に 6 章で結論を述べる。

## 2 実験準備

本論文では当社で独自に収集した、2008 年 1 月 1 日から 2014 年 12 月 31 日までの国内の経済記事約 130 万件を用いた。不祥事記事か否かのラベル付けには社内の不祥事記事リストを用いて経済記事内容と紐付けた。不祥事記事は報道されるすべての記事のうち一部に過ぎず、クラス割合は 1%未満となった。

出力結果に対する簡易に解釈可能であるモデルを構築するため、モデルにはロジスティック回帰モデルを用いた。不均衡データであることを考慮し正のクラスの適合率と再現率の調和平均である F 値を最大化する分類器を学習し、最適化には確率的勾配法を用いた。[2]特徴量には比較的シンプルで解釈性の高い $N$-Gram($N = \{1, 2\}$) を用いた。$N$-Gram 構築のための分かち書き、固有表現認識には mecab[3]を用いた。

本論文では表 1 にある 8 通りのモデルを用いて実験を行ない、それぞれを比較した。経済記事は趣旨を冒頭にまとめることが多いことから、記事の全文を用いた場合、冒頭の 2 文のみ用いた場合を比較した。

特徴調整として記号、数、助動詞や一部の助詞、固有名詞を除いた。

## 3 全モデルの比較に関する実験

最初の実験では学習期間を 2008 年の 1 年間、テスト期間を 2009 年の 1 年間とし、そのときの 8 通りの全モデルでの正解率、適合率、再現率を表 2 に示した。F 値ではモデル 6 が最も良いが、「不祥事記事を見逃したくない」という実務的需要を鑑みれば再現率の高いモデル 1 も注目に値する。

[2]適合率は「不祥事」と予測されたうち実際に「不祥事」の記事だった割合。再現率は実際に「不祥事」の記事のうち「不祥事」と正しく予測された割合。

[3]http://taku910.github.io/mecab/

表 2: それぞれのモデルの精度、適合率、再現率（%）。学習期間は 2008 年 1 月 1 日から 12 月 31 日、テスト期間は 2009 年 1 月 1 日から 2009 年 12 月 31 日。最も値の大きいモデルを**太字**とした。

| モデル | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 正解率（%） | 97.5 | 98.7 | 98.3 | 98.4 | 98.7 | **98.8** | 98.7 | 98.6 |
| 適合率（%） | 12.3 | 18.0 | 16.1 | 15.1 | 19.3 | **20.3** | 20.1 | 18.0 |
| 再現率（%） | **78.1** | 58.9 | 69.2 | 61.3 | 68.3 | 58.8 | 66.1 | 61.3 |
| F 値 | 21.8 | 30.4 | 27.7 | 26.2 | 32.3 | **33.7** | 33.4 | 30.4 |

次にそれぞれのモデルから出力された特徴量の重みベクトルの上位の $N$-gram を表 3 に示した。特徴調整を行わない場合（モデル 1、3、5、7）は「スルガ」「東洋建設」のような固有名詞（二重下線）が結果を左右しがちだが、調整を行うことで、上位に「不祥事に関連する」と解釈可能な下線の特徴が多い。また、表 3b の Bi-gram では、「巨額+損失」や「市況+悪化」のように、単独では解釈しにくい特徴量がより明確になることがわかる。

## 4 特徴調整の重要性に関する実験

なぜ「スルガ」や「東洋建設」のような固有名詞が特徴調整を行わないモデルで重みが大きくなるのか。これを明らかにするために 4 つの実験を行なった。

1 つ目の実験として、モデル 1 の特徴量のうち、それに特徴調整を行なったモデル 2 に含まれないものを調べた。その上位 30 個は、

- スルガ, アスキー, サンエー, 日本たばこ産業, 従来, て, 浦安, アイフル, として, 六, 全日空, 伊藤ハム, 日本製紙, 昨年, 篠原, 二月, 刈羽, モリテックス, ブリヂストン, 三月, によって, 新日軽, ○, 西松建設, テレビ朝日, エーザイ, 長岡, レナウン, 野村, 柏崎, 琉球, 八王子, 任天堂

となっており、固有名詞（下線）が大半を占めており、その中でも企業名が目立つ。

2 つ目の実験（図 1）では不祥事に該当した企業が、1 年以内に別の不祥事記事に再度登場する回数とその頻度等を調べた。1 度も登場しなかった企業が約 25%に過ぎないことから、7 割以上の企業は 1 年以内に別の不祥事記事に出現していることがわかった。

3 つ目の実験（表 4）として、人間の判断により「不祥事に関連すると思われる特徴量」と「不祥事に関連しないと思われる特徴量」が、様々な学習期間（1,3,5 年）でどのように重みが変化するかを比較した。不祥事関連の特徴の場合は年数に関わらず重みが大きいことがわかるが、不祥事に直接関連しない特徴量の場合

表 3: それぞれのモデルから出力された特徴量の重みベクトルの上位 20 個。二重下線の特徴量は固有名詞、数字などを表し、下線は不祥事に関連すると解釈可能な $N$-Gram。

(a) Uni-Gram

| Rank | モデル 1 | モデル 2 | モデル 3 | モデル 4 |
|---|---|---|---|---|
| 1 | スルガ | 問題 | 東洋建設 | 半期 |
| 2 | 社員 | 立ち退く | トラブル | 市況 |
| 3 | ポリエチレン | 行き詰まる | 支払い | 発表 |
| 4 | 重 | 火災 | スルガ | けじめ |
| 5 | 社内 | 採算 | 巨額 | トラブル |
| 6 | 受理 | 当たる | マクドナルド | 偽装 |
| 7 | 疑い | 誤る | 回収 | 障害 |
| 8 | 発表 | 子会社 | 無償 | 有毒 |
| 9 | アスキー | 重 | 長谷 | コーポ |
| 10 | すぎ | モノ | 日航 | 読める |
| 11 | 製品 | ミス | けじめ | 東洋 |
| 12 | サンエー | 巨額 | 製 | いる |
| 13 | 日本たばこ産業 | 取扱 | 伊藤ハム | 不具合 |
| 14 | 停止 | 残業 | で | 軽金属 |
| 15 | 従来 | 不適切 | 社員 | 差す |
| 16 | 障害 | スイスフラン | 半期 | 巨額 |
| 17 | 顧客 | 巡る | 約 | 赤字 |
| 18 | 誤る | 発表 | キリンビバレッジ | 火災 |
| 19 | て | 費用 | 発表 | 操縦 |
| 20 | 無償 | 操縦 | 巡る | 過重 |

(b) Bi-Gram

| Rank | モデル 5 | モデル 6 | モデル 7 | モデル 8 |
|---|---|---|---|---|
| 1 | ギョーザ | よる | リコール | 東洋 |
| 2 | 日本たばこ産業 | 巨額 損失 | 赤字 | 無配 |
| 3 | 子会社 | リコール | 支払い | 障害 |
| 4 | 無償 | いる | 談合 | 偽装 |
| 5 | 同社 | コーポ | 命令 | する |
| 6 | 談合 | 問題 | 日航 | カルテル |
| 7 | 誤る | 損失 | 停止 | トラブル |
| 8 | と 発表 | 赤字 | いる た | 損失 |
| 9 | 伊藤ハム | 行き詰まる | 伊藤ハム | 不具合 |
| 10 | 障害 | 採算 | トラブル | 社員 |
| 11 | ミス | 製 | 子会社 「 | 火災 |
| 12 | 年 三月 | 社員 | 七 年 | マンション 市況 |
| 13 | 三月 期 | 誤る | 無償 | 談合 |
| 14 | スルガ | 製品 | が | 前期 下方 |
| 15 | 支払い | 停止 | 派遣 | 発表 する |
| 16 | 販売 する | 問い合わせ 先 | 火災 | 市況 悪化 |
| 17 | 巨額 損失 | 冷凍 | 年 三月 | 勧告 |
| 18 | が 二 | 障害 | 約 | 円 賠償 |
| 19 | 派遣 | 製造 する | 行き | 建 |
| 20 | 一部 | 同行 | が 二 | 前期 |

は重みのばらつきが見られる。例えば「日航」であれば短い期間でのみ重く、逆に「トヨタ」は長い期間のみ重く、ばらついている。図 1 と合わせて考えると、ある企業の不祥事が報道されると、続けて同じ企業に関する不祥事の報道が続きやすく、これが固有名詞を重視する重みベクトルが学習される原因になっていると考えることができる。

4 つ目の実験（表 5）として、モデル 1 とモデル 2 の F 値、適合率、再現率を異なる訓練期間で比較した。テスト期間は 2014 年で統一し、その直前の 1〜6 年間を学習期間とした。特徴調整をしない場合（モデル 1）では、学習期間の長期化に従って、再現率の著しい低下傾向が見られる。特徴調整をした場合（モデル 2）ではそれほど低下しないことと比較し、モデル 1 では固有名詞による過剰適合が起きている可能性がある。

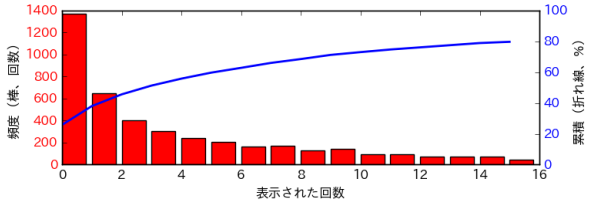


図 1: 該当企業、1 年以内の［不祥事］記事に出現する回数、頻度、累積 (%)。

表 4: 各学習期間ごとの重みを抽出した左表は「不祥事」記事に関連するであろう特徴量、右表は「不祥事」記事に関連しない特徴量。

| 特徴 | 1 年間 | 3 年間 | 5 年間 |
|---|---|---|---|
| 停止 | 23.7 | 24.6 | 24.3 |
| リコール | 23.0 | 21.1 | 26.0 |
| カルテル | 19.7 | 23.9 | 21.6 |
| 遅れる | 18.9 | 21.6 | 20.0 |
| 賠償 | 18.7 | 21.0 | 17.3 |
| 回収 | 17.4 | 18.4 | 16.1 |
| 粉飾 | 16.0 | 10.1 | 8.2 |
| 漏れ | 14.9 | 13.8 | 12.6 |
| 解任 | 14.6 | 12.3 | 11.9 |
| 不具合 | 14.6 | 17.9 | 18.4 |
| 漏れる | 14.1 | 10.8 | 8.8 |
| 損失 | 13.7 | 12.6 | 14.3 |
| 障害 | 12.2 | 15.5 | 14.6 |
| 無配 | 12.1 | 19.2 | 20.3 |
| 事故 | 11.7 | 15.7 | 13.9 |
| トラブル | 11.5 | 17.0 | 15.2 |

| 特徴 | 1 年間 | 3 年間 | 5 年間 |
|---|---|---|---|
| 東電 | 17.0 | 12.0 | 13.8 |
| た | 15.4 | 14.1 | 16.4 |
| オリンパス | 10.2 | 20.6 | 16.3 |
| ドコモ | 8.3 | 8.5 | 7.6 |
| 粉ミルク | 8.1 | 9.9 | 6.1 |
| インデックス | 8.0 | 4.4 | 0.4 |
| 福島 | 7.9 | 5.9 | 4.0 |
| 日航 | 7.7 | 4.2 | 2.0 |
| 検察官 | 7.1 | 4.4 | -0.9 |
| 公取委 | 6.9 | 6.9 | 10.1 |
| 全日空 | 5.7 | 7.5 | 7.9 |
| 日本テレビ | 5.7 | 7.3 | 6.6 |
| 最終 | 5.3 | 9.0 | 9.7 |
| 三菱自動車 | 3.4 | 0.5 | 0.6 |
| トヨタ | 3.0 | 4.7 | 17.0 |
| ゼネコン | 1.9 | 0.9 | 7.7 |

## 5 再学習に関する実験

我々の経済社会では、時間の経過とともに、絶えず新たな種類の不祥事が報道され社会的関心を集めている。そのため高い適合率・再現率を維持するためには、新たなラベル付きデータを追加して分類器を学習し直す必要がある。ただし、データを増やすために毎年 20 万件の経済記事を人手でラベル付けを行うことは、システムを構築した当初の目的に反する。そこでファンドマネジャーが日々、ユーザインターフェース（UI）を通して不祥事関連と予測された記事には、少なくともすべてに目を通していることに着目した。ファンドマネジャーがそれらの不祥事と分類された記事の中で「不祥事ではない」（偽陽性）と感じた記事があれば、UI からラベルを反転してもらい、その記事を分類器の学習データに追加した。これにより適合率の向上が期待できる。再現率の向上のためには、分類器が不祥事ではないと判断した記事について、別の UI を通してそれらの記事の一部だけでも偽陰性のチェックを行うことで、同様にラベルを反転させ再学習に用いた。

本実験では、上記の UI を通したラベル再付与の効果を下記の方法で擬似的に検証した。

表 5: モデル 1 とモデル 2 の F 値、適合率、再現率。テスト期間は 2014 年、学習期間は 2013 年以前の 1〜6 年間とした。

| | モデル 1 | | | モデル 2 | | |
|---|---|---|---|---|---|---|
| | F 値 | 適合率 | 再現率 | F 値 | 適合率 | 再現率 |
| 1 年間 | 29.8 | 18.5% | 77.1% | 35.0 | 23.7% | 67.0% |
| 2 年間 | 40.6 | 30.4% | 61.3% | 36.8 | 25.9% | 63.8% |
| 3 年間 | 43.1 | 38.4% | 49.1% | 38.6 | 28.9% | 57.9% |
| 4 年間 | 40.8 | 43.6% | 38.4% | 40.4 | 30.6% | 59.5% |
| 5 年間 | 36.7 | 44.7% | 31.2% | 38.9 | 30.2% | 54.6% |
| 6 年間 | 27.3 | 45.3% | 19.5% | 31.3 | 22.9% | 49.4% |

表 6: 再学習に用いたモデルの一覧とその学習期間、テスト期間。

| モデル | 訓練データ期間 | テストデータ期間 |
|---|---|---|
| $m_1$ | 2008 年 1 月〜2008 年 6 月 | 2008 年 7 月〜2009 年 6 月 |
| $m_{1,1}$ | 2008 年 1 月〜2008 年 7 月 | 2008 年 8 月〜2009 年 7 月 |
| $m_{1,2}$ | 2008 年 1 月〜2008 年 8 月 | 2008 年 9 月〜2009 年 8 月 |
| ⋮ | | |
| $m_{1,6}$ | 2008 年 1 月〜2008 年 12 月 | 2009 年 1 月〜2009 年 12 月 |
| $m_2$ | 2008 年 2 月〜2008 年 7 月 | 2008 年 8 月〜2009 年 7 月 |
| $m_{2,1}$ | 2008 年 2 月〜2008 年 8 月 | 2008 年 9 月〜2009 年 8 月 |
| ⋮ | | |
| $m_{67,6}$ | 2013 年 1 月〜2013 年 12 月 | 2014 年 1 月〜2014 年 12 月 |

1. 2008 年 6 月から 2013 年 12 月まで月次で過去 6ヶ月のデータをもとに $m_i$ $(i = 1, 2, \ldots, 67)$ を学習
2. 各 $m_i$ につき、翌月の「不祥事」記事を予測する
3. 予測結果と正解データを比較し、予測結果を修正する[4]
4. 修正したデータを追加し、$m_i$ を再学習させ、$m_{i,1}$ を得る
5. 2. から 4. を 6ヶ月分繰り返し、$m_{i,j}(i = \{1, 2, \ldots, 67\},\ j = \{1, 2, \ldots, 6\})$ を得る
6. 修正による効果を検証するため、表 6 のように $m_i, m_{i,j}$ に対し学習終了時点から 6ヶ月先までの「不祥事」記事予測を行う

結果の適合率、再現率をそれぞれ図 2 の左右に示した。$Lag = 1$ の場合、毎月モデルの出力結果を修正した場合の適合率および再現率の中央値を示した。$Lag = 2, ..., 12$ の場合、モデルの出力結果を $Lag$ の月数だけ修正せずにモデルの学習終了時点 $+Lag$ から 12ヶ月先までを予測した時の適合率および再現率の中央値を示した。この結果から、適合率・再現率ともにデータを修正しない月が経過するほど影響を受けやすくなることが観察できる。

## 6 終わりに

**まとめ** 経済記事を「不祥事」または「不祥事ではない」に分類するためのシステムを設計・構築し、再学習の仕組みも取り入れた。実運用においては精度は重要指標の一つに過ぎず、解釈性・頑健性の向上を重視した。

[4]偽陽性は全ての正を負に修正し、偽陰性は 5 割の確率で負を正に修正した。偽陽性は発見が容易であるが、不祥事ではない記事は必ずしもそうではない UI 上の性質を反映させた。

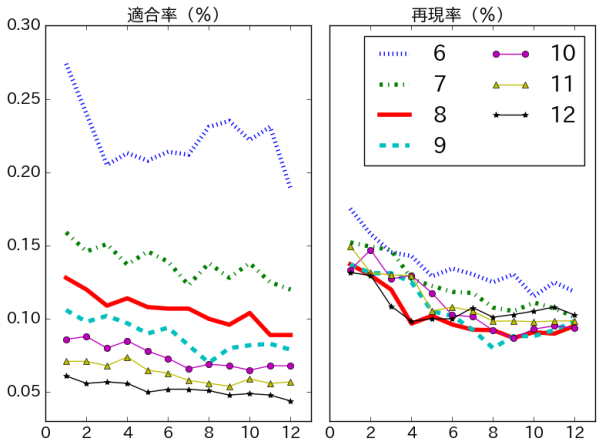


図 2: モデルを更新しない場合の影響。左図は適合率の中央値、右図は再現率の中央値（%）を縦軸に示し、横軸は Lag。凡例は学習期間の月数。

**今後の課題** 予備的分析より実際には不祥事記事だが、「不祥事」というラベルがつけられていないケースがあることがわかった。逆に「不祥事」というラベルがついているのに、不祥事ではない記事は確認できなかった。この場合、「不祥事」というラベルが付いている記事以外を「不祥事ではない」（負例）として扱うより、「ラベル無し」として扱う方が自然であり、（クラス割合の不均衡な状況での）正例とラベル無しデータからの学習手法 [2] の利用を検討したい。

## 参考文献


[1] M.T. Ribeiro, S. Singh, and C. Guestrin. Model-agnostic interpretability of machine learning. In *ICML Workshop on Human Interpretability in Machine Learning*, 2016.

[2] T. Sakai, M. C. du Plessis, G. Niu, and M. Sugiyama. Semi-supervised AUC optimization based on positive-unlabeled learning. In *Machine Learning*, 2017.

[3] F. Z. Xing, E. Cambria, and R. E. Welsch. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 2017.