PatchWorks

【田中頼人特別研究】

第6回レポート

2301330039：安田直也

中川 慧、平野 正徳、藤本 悠吾：大規模言語モデルを活用した金融センチメント分析における企業固有バイアスの評価、第21回テキストアナリティクス・シンポジウム、vol.124, no.173, NLC2024-15、pp.81-86（2024年9月3日）

企業固有のバイアスの定量化:

企業名をプロンプトに含めた場合と含めない場合で、LLMが生成するセンチメントスコア（感情スコア）を比較し、企業固有のバイアス（差分）を定量化しています。これにより、LLMが特定の企業に対して偏ったセンチメントを示すかを評価しています。

企業固有バイアスの企業特性との関連性評価:

企業固有バイアスがある企業の特性（例：規模、モメンタム、バリュー）を分析し、どのような企業がバイアスを受けやすいかを特定しています。これにより、LLMのセンチメント分析におけるバイアスがどのような特徴に影響を受けるかを明らかにしています。

株価パフォーマンスへの影響の理論的および実証的評価:

経済モデルを構築（バイアスのある投資家、ない投資家が一定の割合と定義してリスク資産の価格変動を定義し市場均衡価格を導出、etc）し、偏ったLLMの利用が投資家行動や市場価格に与える影響を理論的に解析しています。また、日本の実際の財務データを用いて、企業固有バイアスが株価に与える実証的な影響（異常収益など）を調査しています。

複数のLLMモデルによる比較:

GPT-4やGPT-3.5などの異なるLLMを使用し、各モデルのバイアス特性を比較しています。これにより、モデル間でバイアスの傾向や程度がどのように異なるかを示しています。

バイアスのある企業とない企業の株価の累積異常収益（CAR）を比較し、バイアスが株式市場でどのような影響を及ぼしているかを時間軸で評価した。

Kamruzzaman, M.; Nguyen, H. M.; Kim, G. L.: "Global is Good, Local is Bad?": Understanding Brand Bias in LLMs, arXiv preprint, arXiv:2406.13997, (2024)

グローバルブランドとローカルブランドに対する属性の関連付け:

方法: 「Stimulus to Attribute Inference (SAI)」および「Attribute to Stimulus Association (ASA)」の2方向でブランドと属性の関連付けを評価。LLMはグローバルブランドをポジティブな属性、ローカルブランドをネガティブな属性に関連付ける傾向を確認した。特にGPT-4oが最も強いバイアスを示した。

SAI: ブランド名を提示し、そのブランドに関連する属性（ポジティブ、ネガティブ、ニュートラル）をLLMが選択。

ASA: 属性を提示し、その属性に関連するブランドをLLMが選択。

経済的バイアスの評価:

方法: 高所得国および低所得国の個人向けのギフトとして、LLMがラグジュアリーブランドと非ラグジュアリーブランドをどのように推薦するかを調査。高所得国向けにはラグジュアリーブランド、低所得国向けには非ラグジュアリーブランドを一貫して推薦する傾向が確認。Gemma-7Bは高所得国向けに100%ラグジュアリーブランドを推薦。

国別ローカルブランドの嗜好の定量化:

方法: 「出身国（Country of Origin）」を指定した場合に、LLMがローカルブランドを選好するかどうかを調査。出身国が指定された場合、多くのモデルがローカルブランドを好む傾向を示した。

LLMのバイアスが製品推薦や市場分析において、グローバルブランドを優遇し、ローカルブランドを不利にする可能性を示した。

Quiñonero-Candela, J.; Wu, Y.; Hsu, B.; Jain, S.; Ramos, J.; Adams, J.; Hallman, R.; Basu, K.: Disentangling and Operationalizing AI Fairness at LinkedIn, Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), vol. 124, no. 173, pp. 81-86 (2023).

測定とギャップの特定:

PAV（People Also Viewed）モデルで予測パリティギャップ（例えば、男性と女性間のスコア差）を検出。

性別による予測スコアが実世界の結果と一致しない場合の偏りを測定しました。

性別の偏りがどのように発生したのかを特定：

データ分布、特徴の妥当性、エラー分析を実施。

緩和戦略の評価：

性別情報を使用せずにギャップを解消する方法を試したが成功せず、最後の手段として性別を用いた後処理を採用。

この後処理により、ギャップを解消し、モデルの公平性が向上した。

バイアス解消の評価：

性別スコア調整後の影響を、接続リクエスト受け入れ率、メッセージ返信率、問題報告率などのメトリクスで評価。

性別に関するケーススタディにより、LinkedInのフレームワークが公平性向上に有効であることを示し、モデルの全体的なパフォーマンスも改善されることを示した。

研究テーマのタイトル：

AI検索サービスにおける企業優遇バイアス：市場競争への潜在的リスク

先行研究との違い：

①AI検索サービスに焦点を当てていること：

先行研究が汎用的な言語モデルに焦点を当てているのに対し、本研究ではPerplexity、Genspark、Google AI OverviewsなどのAI検索サービスを対象にし、実際の検索結果の挙動を分析します。

②企業優遇バイアスに焦点を当て競争市場への影響を調査すること：

多くの先行研究が社会的・文化的バイアス（例: 性別、民族、文化的ステレオタイプ）に注目しているのに対し、本研究では企業優遇バイアスに焦点を当てます。また、このバイアスが市場競争の公正性に与える潜在的なリスクについて考察します。

研究テーマのタイトル：

AI検索サービスにおける企業優遇バイアス：市場競争への潜在的リスク

評価の方針：

①プロンプトを用いたバイアスの定量的評価：

先行研究（1）を参考にし、企業名を含むプロンプトと、マスクしたプロンプトの二種類をAI検索サービスに入力して1～5の感情スコアを得る。その差分を企業優遇バイアスの定量的指標として評価し、サービス間で比較を行います。

②アンケートによる印象変化の測定：

具体例として、クラウドサービスを推薦するプロンプトを使用し、AI検索結果を閲覧する前後で、利用者の印象がどのように変化したかをアンケートで収集します。これにより、AI検索結果が特定の企業やサービスに対する認識に与える影響を定量化します。

③市場への潜在的リスクの考察：

①と②の結果を基に、特定企業やサービスカテゴリーがバイアスを受ける割合を算出します。その結果をもとに、AI検索サービスの普及が進んだ際の市場競争への影響を定義し、潜在的なリスクを議論します。この議論は、競争法や市場公正性に関する政策提言にも繋げます。