

- 分析者と記者とのコミュニケーションの中で探索的に分析観点を見つけるための手法を提案する
- 文埋め込みに対して独立成分分析を実行した結果からカテゴリ変数を作り、それを特徴量として分析に利用する手法を提案した
- 実データに対して提案手法を利用したケーススタディを通して、その有用性を検討した

モチベーションと概要

データジャーナリズムの重要性の高まりに伴い、報道現場では取材の仮説構築のためのテキスト分析が求められている。特に現場ではテキストのドメインへの深い知識を持っていない場合も多い。そのような場合において分析者と記者の間でコミュニケーションをしつつ、探索的に分析観点を見つけていくための手法を検討し、得られた分析結果（図1, 2）とその示唆をケーススタディとして述べる。

カテゴリ1	消費税増税の影響と連休の影響による小売・不動産・飲食業界の売上低迷
カテゴリ4	商店街の物販低迷と飲食業界の過当競争による地域経済の冷え込み
カテゴリ6	宿泊業界における外国人観光客減少と連休効果の限界
カテゴリ15	製造業における受注・販売減少と構造的課題
カテゴリ24	暖冬影響による冬物商材の売れ行き悪化と 年末年始の消費低迷
カテゴリ30	消費税増税と米中貿易摩擦による国内外の消費と観光業界の不振

図1. 得られたカテゴリの名前（GPT-4によって命名）

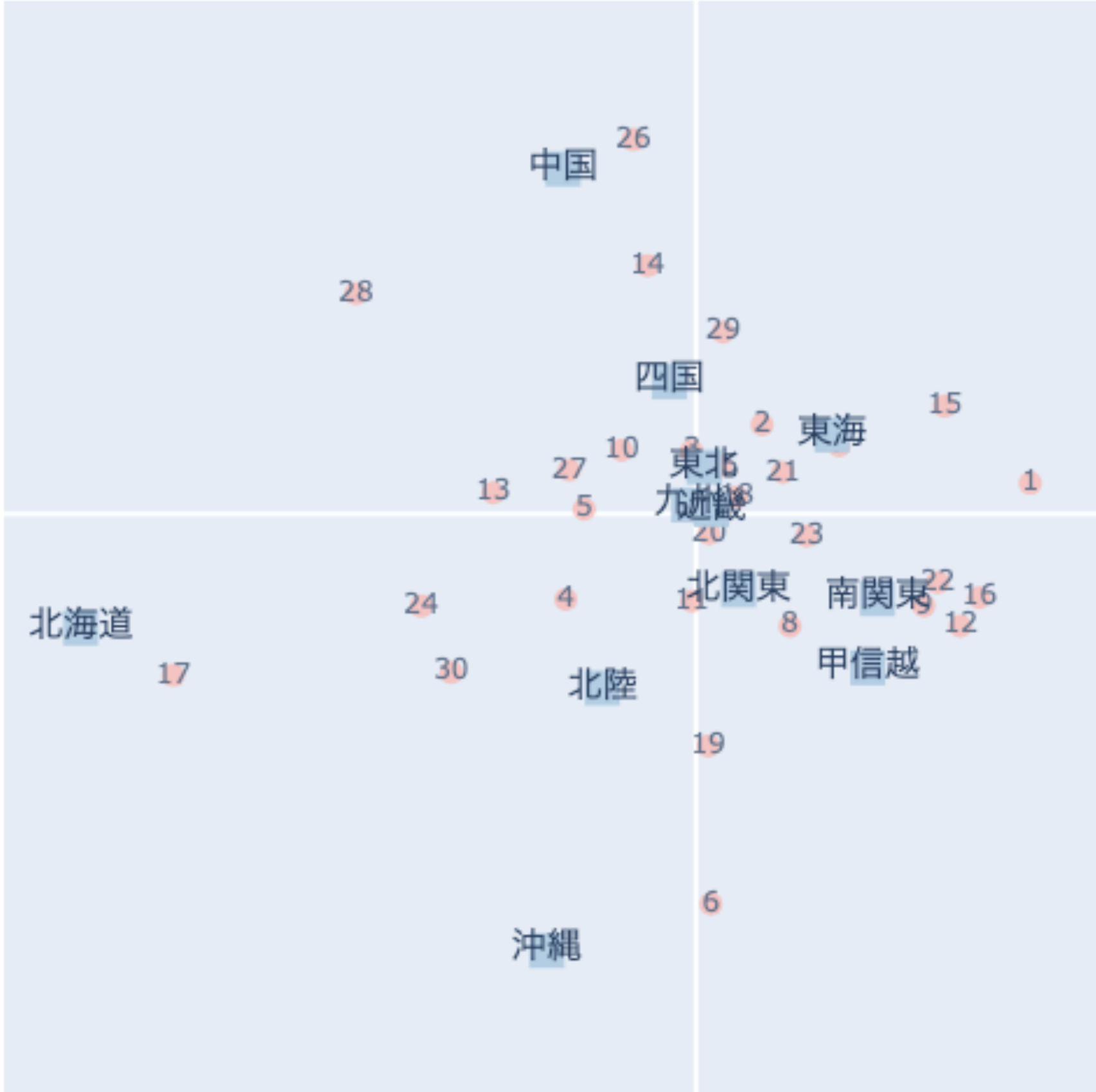


図2. カテゴリ化された独立成分を対応分析した結果

分析手法

本研究の分析手法は3つの工程から構成されている。

- ① テキストを文埋め込みに変換する
- ② 独立成分分析（ICA）によって文埋め込みを次元削減する
- ③ 文の独立成分スコアを使ってカテゴリ変数に割り当てていく

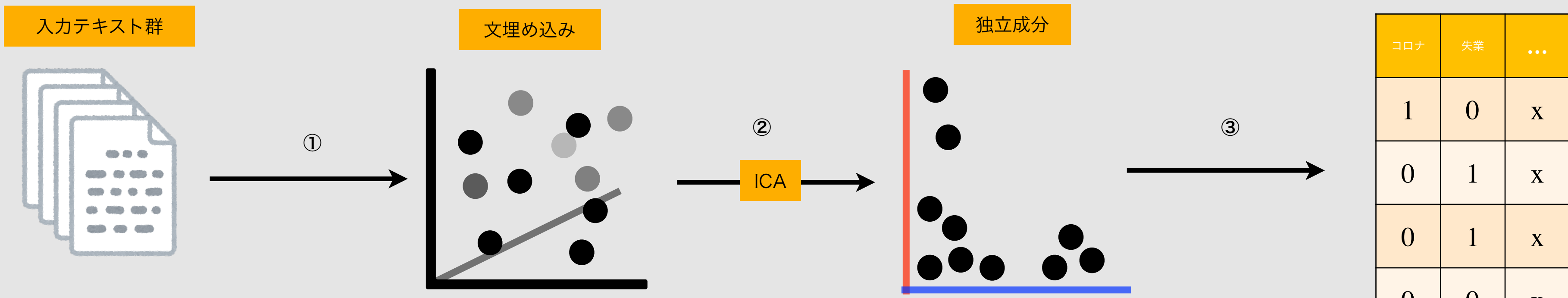


図3. 提案手法の概要図

データセット

内閣府の景気ウォッチャー調査に本手法を適用し、ケーススタディとした。このデータは地域・業種別の景気の現状と先行きに対する評価とその判断理由のテキストからなる。2019年1月から2022年12月までの景気の現状について悪いとやや悪いを選択した9,070件の判断理由を分析をした。

- 今シーズンに入り来店客の減少、客単価の減少が顕著となり、アパレルを筆頭に冬物商材の動きが鈍く、ほとんどのテナントが前年実績未達成の状況である
- 新聞へ出稿する求人広告が一段と減少した。広告会社も新聞媒体は売れないといっている。
- 前年に比べ年末商戦は悪いとの報告が多い。

図4. 景気判断理由のテキストの実例（抜粋）

ケーススタディ

▶景気ウォッチャーを対象に「記事にできそうな話題の種を見つけられるか？」という要求に対応する

テキストの特徴から短文に対して有効であり、軸に対して意味を解釈できるという解釈の容易性の観点で文埋め込みに対する行列分解をベースとした本手法を選んだ。さらに、行列分解の手法は説明可能性が高いと先行研究[1]によって示唆されている点でICAを利用した。その結果の一部が図5であるが、例えば独立成分1のスコア上位には、サービス業の売上低迷に関するテキストが集まっており、直感的な結果が得られた。

得られた分析結果とさらなる応用例

提案手法によってテキストをカテゴリ化(*1)した集計結果が表1である。

カテゴリの解釈のために、独立成分のスコアが高いテキストをプロンプトに加えて、GPT-4にカテゴリ名を命名させた。図1はその結果の一部の抜粋である。

さらなる分析のため、本手法を適用後に、カテゴリに属するテキストの数を地域ごとに集計し、この結果を目視でざっと確認できるように多重対応分析を実行した結果が図2である。

結果の解釈例とその活用方法

一例として本結果から カテゴリ24 = 暖冬による冬物商材の影響 が北海道や北陸が他の地域よりも強い

（図2）ということが読み取れる。暖冬という単語を含んだテキストは9,070件の中の51件と全体の約0.005%の割合しかなく、このようなカテゴリのまとまりは単語のカウントベースの手法で発見することは非常に難しく、本手法の有用性が現れていると考えられる。

このような示唆を得ることができれば、「豪雪地帯では暖冬の影響が強いのではないか？」という仮説を立てられる。そして、この結果からPOSデータなどの別の統計データに当たったり、現地に取材しに行くといった判断ができる。

[1] Yamagiwa et. al , 2023, Discovering universal geometry in embeddings, EMNLP 2023

報道現場において発生するテキスト分析の需要に対して、その要求に応える分析手法を提案した。さらに、その手法をケーススタディを通して、実際のデータに対して分析での実用例を確認することで、本手法の有用性を確認した。一方で、本研究では他手法に対する優位性などは評価できておらず、この観点における評価手法の確立等を今後の課題としたい。

まとめ

*1 文埋め込みには塚越ら[2]のcl-nagoya/sup-simcse-ja-largeを利用し、ICAの成分数に30を指定している

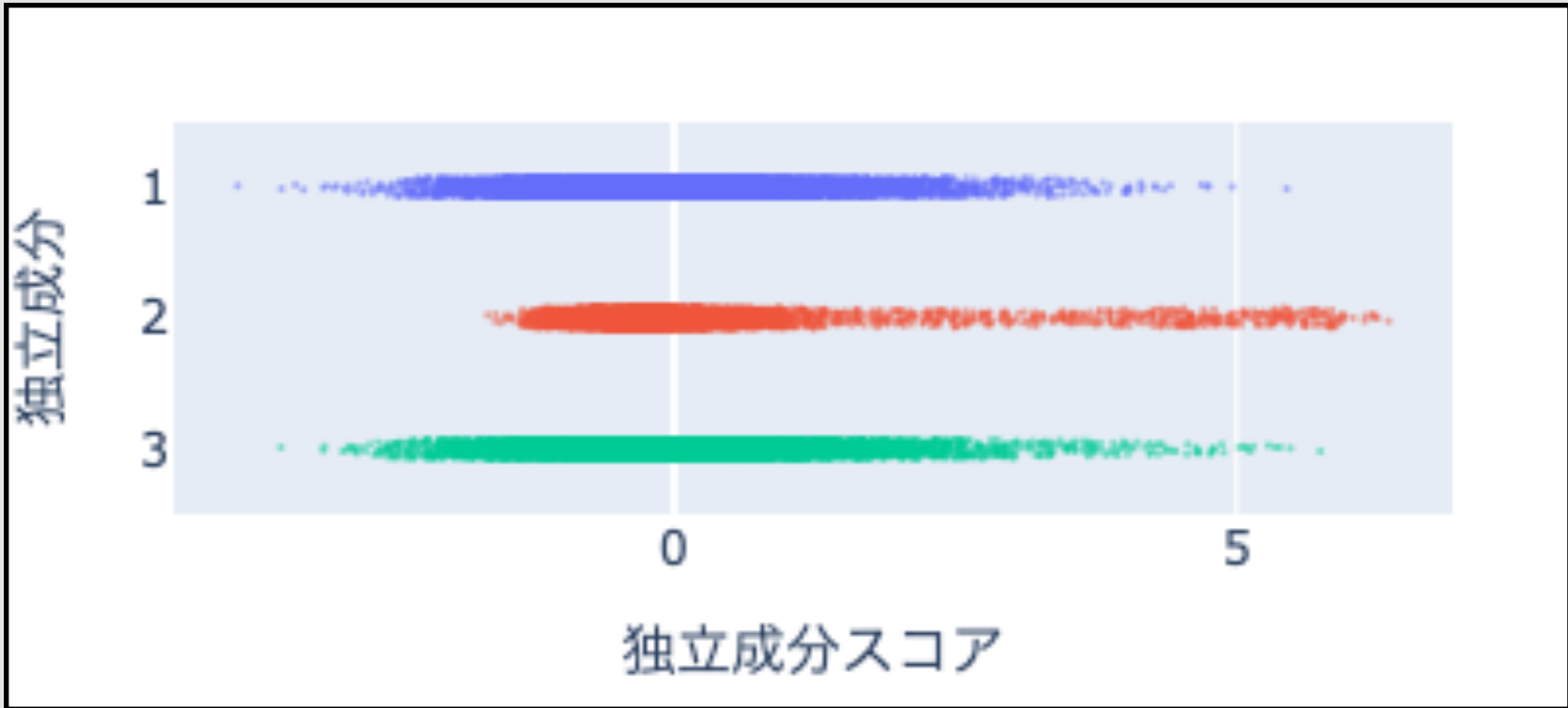


図5. 3つの独立成分とそのスコアのストリッププロット

カテゴリ	1	4	6	15	24	30
北海道	8	40	15	11	50	67
東北	21	43	14	34	50	30
南関東	91	70	56	130	48	73
北関東	23	27	28	28	17	16
東海	85	36	37	73	30	76
甲信越	27	26	21	19	21	20
北陸	11	27	42	11	42	30
近畿	26	33	47	50	51	81
中国	31	38	15	46	29	40
四国	6	20	11	36	14	22
九州	23	43	33	37	34	47
沖縄	3	12	28	3	2	33

表1. カテゴリとそれに属する文の集計結果（抜粋）