# RoBERTa を用いた経済不確実性のテキスト分類

桑名祥平　佐々木稔
茨城大学　工学部　情報工学科
19t4026r@vc.ibaraki.ac.jp　minoru.sasaki.01@vc.ibaraki.ac.jp

## 概要

近年，新聞記事や SNS などに含まれる政治や経済に関連するテキストを対象として様々な分析を行う研究が盛んにおこなわれている。本論文では経済分野の中で、経済指標について焦点を当てた。GDP や消費者物価指数といった社会経済を表す指標が多数存在するなかで、テキストデータを用いた指標も注目され始めている。その中で、”経済政策不確実性指数”と呼ばれる現状の社会経済が持つ不確実な要因を数値化した指標に着目し、入力したテキストデータに対して機械学習手法を用いて不確実性を持つかどうかを分類する手法を提案する。提案手法は RoBERTa を用いて二値分類問題を解くもので、不確実性を持つテキストと持たないテキストにラベル付けされた訓練データで分類モデルを学習し、得られた分類モデルにテキストを入力して不確実性を持つかどうか分類する。ラベル間のテキストの類似度も考慮したデータセットを用いることで、モデル性能の向上を確認した。

## 1 はじめに

近年，デバイスの小型化や大容量化、通信の高速化などにより、あらゆる分野においてデータを活用する動きが活発である。従来は人手で行われていた活動もデータを用いた自動処理に置き換わることも多くある。中でも近年着目されている技術が、経済分野を対象としたテキストデータの分析である。
日本には国内総生産や景気動向指数、企業物価指数などの経済指標があり、これらの指標は莫大な時間とコストをかけて算出されている。そのような指標の中で、大量の新聞記事のテキストデータを用いた経済指標のひとつとして”経済政策不確実性指数”というものが存在する。この指数の算出に用いられるテキストデータは、経済に関する新聞の中でも、不確実性に言及している記事のみであり、人手で抽出するには専門的な知識と経験が必要である。そこで、本稿では、経済政策不確実性の算出に用いられる新聞記事を抽出するために、入力された記事に対して不確実性についての記述があるかどうかを自動的に判別する手法を提案する。具体的には、経済に関する新聞記事のテキストデータの中で、不確実性に言及されているか否かをラベルで表現し、RoBERTa を用いて二値分類問題として分類モデルを学習し、入力テキストに対してこの分類モデルを用いて不確実性の言及があるかどうかを分類する。実験では、データセットの各文章をベクトル化し、ラベル間の類似度によって評価データに対する判別の正解率がどのように変化するのか、また適切なデータセットはどのようなものなのかを調査した。ラベル間の類似度に関する説明は 3 章 2 節を参照されたい。

## 2 関連研究・関連手法

テキストデータから経済指標を算出する研究は盛んにおこなわれている。景気ウォッチャー調査と呼ばれる日本各地でとられた経済に関する 5 段階の評価と選択理由を学習データ、テストデータに日経新聞を用いたセンチメント分析による指標を構築した研究[1]や、日本銀行が発行するテキストに対してトピックモデルやニューラルネットワークを用いて指標を構築した研究[2]、さらに本稿でも扱う、新聞記事に含まれる不確実性について指数化した経済政策不確実性指数 [3]などさまざまである。

本稿の先行研究[4]では、経済政策不確実性指数の算出に用いられる不確実性を含むテキストの抽出を行っている。具体的には、経済に関する新聞記事のデータセットを用意し、不確実性に言及があるものはラベル 1、それ以外を 0 として、SVM や Bidirectional RNN、BERT, DistilBERT, RoBERTa など複数のモデルで学習し、比較を行っている。F 値が最も高かったモデルは RoBERTa であったため、本稿でもこのモデルを採用している。

# 3 提案手法

3 節では，新聞記事のテキストデータから、不確実性に言及されている文章を抽出する方法を紹介する。

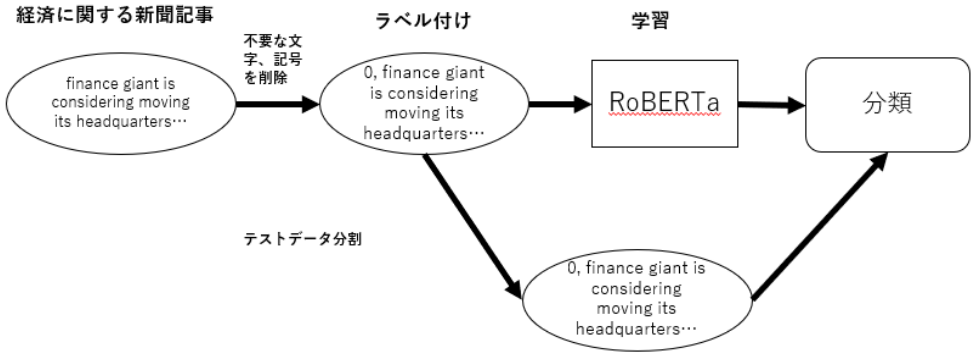


**図 1 ：提案手法の概略**

## 3.1 概要

図 1 に提案手法の概略図を示す。前段階として、経済に関する英字の新聞記事のデータセットを作成している。そして前処理を行い、データセットのうち 80%を学習データ、20%をテストデータと検証データとしている。機械学習では、BERT の派生である RoBERTa を採用し、ファインチューニングを行っている。

## 3.2 データの前処理方法

自然言語処理や画像処理の分野では、ノイズや欠損、エラー値などが存在するため、前処理を行うことが一般的である。今回の処理の中でも前処理を行っている。具体的には、経済に関する新聞記事のデータセットを用意し、大文字小文字の処理や省略文字、不必要な記号の削除等を行っている。次に新聞記事に含まれる不確実性のラベリングである。テキストの中にキーワードとして”uncertain”, “uncertainty”、または同様の意味である”unclear”, “unpredictable"などの単語が含まれている場合にはラベルを 1, それ以外のテキストには 0 をラベル付けしている。テストデータ、検証データに関して、ラベルが 1 のテキストに含まれる上記の単語群は省く処理をしている。

上述のように前処理としてラベリングを行ったが、経済政策不確実性に言及があるテキストは、記事全体のうちおよそ 5%ほどであった。学習の際にはラベル間の比率は 1:1 で行うため、学習に使用しないラベル 0 のデータが大量に余る。そこで、Google News データセットの学習済み単語ベクトル用いて各文章に対して 300 次元のベクトル化を行った。1 文の単語の集合を$S = \{w_1, w_2, \dots w_n\}$とすると、1 文章のベクトル$V$の和は以下の式で表せる。

$$V = \sum_{w_i \in S} v_{w_i} \tag{1}$$

ここで 1 文のベクトル$V$は 300 次元のベクトルである。$V = \{x_1, x_2, \dots x_n\}$とし、(1)の式を正規化すると以下の式で表せる。

$$Vs = \frac{V}{|V|} \tag{2}$$

$$|V| = \sqrt{\sum_{k=1}^{300} x_k^2} \tag{3}$$

ラベル 1 のテキストのベクトルの代表値を作り、各ラベル 0 のベクトルとコサイン類似度を取る。ラベル 1 のテキストのベクトルの集合を$V = \{v_{11}, v_{12}, \dots v_{1n}\}$とすると、ラベル 1 のベクトルの総和$V_1$は

$$V_1 = \sum_{v_{1i} \in V} v_{1i} \tag{4}$$

(2)，(3)同様に総和を正規化するとラベル 1 のベクトルの代表値が得られる。

$$V = \frac{V_1}{|V_1|} \tag{5}$$

$$|V_1| = \sqrt{\sum_{k=1}^{300} x_k^2} \tag{6}$$

(5)でえられたラベル 1 の代表値$V$と、各ラベル 0 のテキストに割り振られた(2)式の$V_s$のコサイン類似度を取ると、あるラベル 0 のテキストとラベル 1 の代表値の類似度は以下の式で表せる。

$$\cos_{V_1 V} = \frac{V_s \cdot V}{|V_s||V|} \tag{7}$$

(7)で得られた値がラベル間の類似度である。

### 3.3 エンコード手法

先行研究の結果から、ラベル付きのテキストデータから分散表現を求めるためのエンコード手法として RoBERTa を用いる。事前学習済みモデルは roberta-base を採用している。学習時の各パラメータを以下の表 1 に示す。

**表 1, RoBERTa のパラメータ**

| パラメータ | 値 |
|---|---|
| 学習バッチサイズ | 8 |
| 評価バッチサイズ | 8 |
| エポック数 | 4 |
| 最大シーケンス | 200 |
| 学習率 | $4.0\times10^{-4}$ |

## 4 実験

### 4.1 データセット

本実験で使用するデータセットは、マイクロソフトが提供するマイクロソフトニュースに関するデータセットである MIND[1]に加え、ロイター通信の過去の新聞記事の二つのテキストデータにラベリング、類似度の振り分けを行ったものである。[2]二つのテキストデータではあらかじめ経済に関するテキストのみを抽出しており、約 11000 件のデータセットとなっている。データセットの内訳は、ラベル 1 が 476 件、ラベル 0 のデータが 10740 件となっている。ラベル 0 のデータセットにおける類似度別の内訳を以下の表 2 に示す。本実験では、476 件のラベル 1 のデータ 476 件に対して、約 11000 件のラベル 0 のデータから類似度別に 476 件を抽出し学習、評価を行う。具体的には、ラベル 1 との類似度が 0.9 以上のラベル 0 のテキストのみを抽出し学習するといった流れである。類似度が 0.7 以下のデータセットは、ラベル 1 の件数よりも下回るため、対象外としている。

**表 2. ラベル 0 の類似度別件数**

| 類似度 | 件数 |
|---|---|
| 0.9~1.0 | 4735 |
| 0.8~0.9 | 5185 |
| 0.7~0.8 | 605 |
| 0.6~0.7 | 122 |
| ~0.6 | 93 |

### 4.2 評価方法

評価の方法としては、データセットを分割し、複数回実行する K 分割交差検証を行う。特に今回は分割するデータセットのラベルの偏りを無くすために層化 K 分割交差検証で評価を行う。今回の実験では分割数を 5 に設定しているため、データセットを 5 分割し 80%を学習データ、20%をテストデータとして 5 回実行し、F 値の平均値を取り評価を行う。ラベルが 0 のデータに関して、データセットから無作為に抽出したデータセット「類似度高」、「類似度低」、「高低 5:5」，「高低 3:7」，「高低 7:3」，「類似度 0.9 以上」，「類似度 0.8 以上 0.9 未満」，「類似度 0.7 以上 0.8 未満」である。ここで類似度高、類似度低は、データセットを類似度でソートし

[1] MIND データセット https://msnews.github.io/

[2] ロイターニュースデータセット https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection

たときの上位と下位という意味である。

# 5 実験結果

4.2 で紹介した類似度別の 8 パターンについて、f 値による評価の実験結果を以下の表 3 に示す。

表 3. 各 f1 値

| 類似度 | F 値 平均値 |
|---|---|
| 先行研究 | 0.45 |
| 無作為抽出 | 0.74 |
| 高 | 0.42 |
| 低 | **0.96** |
| 高低 5:5 | 0.79 |
| 高低 3:7 | 0.74 |
| 高低 7:3 | 0.83 |
| 90% | 0.52 |
| 80% | 0.81 |
| 70% | 0.94 |

## 5.1 先行研究との差異

先行研究では、ラベル 1, ラベル 0 の比率が 1:4 で行われており、最も高い F 値を出した RoBERTa で 0.45 であった。本実験では、ラベル間の類似度を考慮しない実験で F 値 0.74、ラベル 0 に関して類似度の低いデータを抽出し学習した場合には F 値は 0.96 と先行研究よりも高いスコアを確認した。

## 5.2 類似度による影響

今回の実験ではラベル比率を 1:1 で固定し、データセットのテキストの類似度に着目して学習を行った。結果、類似度が低いテキストを多く取り入れると F 値は無作為抽出よりも 0.2 ほど上昇した。反対に、似度が高いテキストを多く取り入れると f 値は 0.3 ほど低い値を示した。類似度のパーセンテージで見た際に、ラベル 1 とラベル 0 の類似度が 90%を境に F 値が大幅に変化していることが見て取れる。この結果は、ラベル間の類似度が高い場合、分類の際の境界が曖昧になり正しく分類できず、反対に類似度が低い場合には境界が明確であるから正しく分類できると予想できる。

# 6 終わりに

本稿では、経済政策不確実性指数の算出に用いられる不確実性に言及のあるテキストを RoBERT のファインチューニングを用いた二値分類問題での抽出を行った。実験では先行研究での F 値を上回る結果となった。先行研究の F 値を上回った理由は、ラベルの比率を均一にしたこと、データセットを類似度別で作り学習したことなどが考えられる。

# 参考文献


[1] K. Seki et al. (2022). “News-based business sentiment and its properties as an economic index”. SienceDirect.

[2] K. Yono., & K. Izumi. (2017). “Real time sentiment analysis of Bank of Japan using text of Financial report and macroeconomic index”. The 31st annual Conference of the Japanese Society for Artificial Intelligence.

[3] R.Baker et al.(2016). “Measuring economic policy uncertainty”. The Quartwrly journal of economics. Vol.131. issue 4.

[4] Godbole. S. et al.(2020, 7, February). “Economic uncertainty Identification Using Transformers - Improving Current Methods”. Seminar Information Systems(WS19/20).