テレビ番組の放送内容テキストを用いた視聴者属性別の視聴量変動の予測

山田祐也 [1,2]　南條浩輝 [3]
[1] 滋賀大学大学院データサイエンス研究科
[2] テレビ愛知株式会社　[3] 滋賀大学データサイエンス学部
[1]s6022145@st.shiga-u.ac.jp　[3]hiroaki-nanjo@biwako.shiga-u.ac.jp

概要

近年，テレビ視聴データの利活用が進められており，番組の視聴分析や CM 効果測定などの取り組みが行われている．本稿ではテレビ番組のコーナーごとに視聴量が増加するのか，そうでないのかの予測に取り組んだ．予測には番組内容に関するテキストを用い，BERT モデルを基準に 2 つの精度改善アプローチに取り組んだ．予測精度は視聴者属性によるばらつきがあったものの高い精度で予測できる視聴者属性があった．予測モデルの方法と結果を紹介する．

1　はじめに

1.1　研究背景

近年，視聴データを活用した取り組みが拡大している．視聴データとは，インターネット接続のあるテレビ受像機から収集できるデータである．これを用いることで，いつどの番組を視聴していたかが分かる．しかし放送局は自局の視聴データしか取得出来ない．そこで，在京局，在阪局や在名局などが同一エリアでの視聴データの共同利用を目指した実証実験が進められている [1]．視聴データの活用を目指す背景には，テレビの価値を示すためのビッグデータとしての期待があり，広告代理店やテレビ CM の出稿を検討している企業からはマーケティングデータとしての活用が期待されている．

1.2　研究目的

本研究の目的は次の 2 点である．1 つ目は蓄積されるデータの利活用方法についての知見を得ることである．1.1 節の研究背景で述べたように視聴データの活用方法を考案する．2 つ目は視聴者属性ごとに視聴量変動を予測することである．テレビ番組では，番組コーナー間にテレビ CM を放送している．放送回により視聴量の変動は同一ではなく，番組の後半に視聴量が増大するパターンや減少するパターンなど様々である．視聴量が増加した後の CM チャンスに CM 数を増やすことができればより多くの視聴者にテレビ CM を見てもらうことが出来る（図 1）．そのために番組のコーナーごとに視聴量変動を視聴者属性別に予測することは重要である．本研究では実際に放送による検証は行わないため，予測結果に応じて CM チャンスの CM 数を変動させた場合には視聴行動の変化が起きると考えられるが，これについては本研究では言及しない．

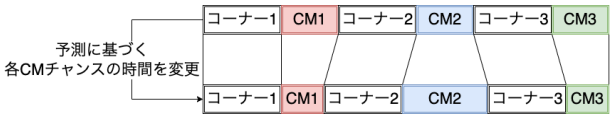

図 1　CM チャンスの運用方法

2　関連研究

本研究で使用する視聴データを用いた研究事例をまとめる．視聴データを用いた研究には放送局とテレビ機器メーカーの研究がある．放送局の研究では同一エリアの複数の放送局が視聴データを連携させる技術検証がある [1]．その中で在名局の実証実験では 5 局の視聴データ，動画配信サービスの視聴履歴とインターネット利用履歴を突合させる検証が行われた．この実験では 5 局の視聴データの連携のみならず，インターネットの検索行動などが紐づけられている．これによりテレビとインターネットをクロスデバイスで分析可能になる．

テレビ機器メーカーの研究には菊池らの研究 [2] があり，番組ジャンルによる視聴傾向の違いを明らかにし，視聴傾向をもとに視聴者属性を推定する手法が提案されている．また，水岡らの研究 [3] では

視聴パターンをクラスタリングする手法が提案されている．

これまでの視聴データに関する研究は共同利用のための技術検証や 1 週間単位の視聴パターンに着目したものが多い．本研究の特徴は視聴データと放送内容に関するメタデータを結びつけ，放送内容に基づく視聴傾向を分析する点である．

3　データ概要

本研究では，視聴データを TVS REGZA 株式会社 [4]，放送内容テキストとして番組シーンメタを株式会社エム・データ [5] から提供していただき研究を行なっている．研究対象の番組は非公開とする．

3.1　視聴データ

視聴データとは，インターネット接続のあるテレビ受像機から取得できる番組視聴データである．本視聴データでは REGZA というテレビ機器から取得されたデータを用いる．データは以下のような特徴があり，放送局や広告代理店などのテレビ業界では広く導入されているデータである．

- 大規模なサンプルで中京エリアでは約 27 万台 (2023 年 11 月末時点)
- １秒単位で視聴量が取得されており詳細に分析が可能
- 収集データは個人利用のテレビに限定されて性別/年代情報が正確

データ項目は番組開始から終了まで１秒単位で視聴者属性別に視聴量が収集されている．本研究ではライブ視聴の視聴量を用いる．放送時刻に対象番組を見ていた割合である．また視聴者属性区分は表 1 の通りである．

表 1　視聴者属性

	20-34 歳	35-49 歳	50-64 歳	65 歳以上
男性	M1	M2	M3	M4
女性	F1	F2	F3	F4

3.2　番組シーンメタ

番組シーンメタは番組のコーナー (話題) ごとに放送時間やその内容を要約したテキストデータからなるデータである．要約の粒度は放送内容を見なくても理解できるほど詳細に制作されており，取材された企業名，商品名や施設名などの情報が収録されている．テキスト中には【施設】「東京駅」，【商品】「ビール」のように番組内で取り扱われた施設名，商品名や【ゲスト】「人物名」や【出演】「人物名」のように出演者情報についても正確に記されている．このコーナー単位で視聴量が視聴者属性別にどのように変動するかを予測する．

3.3　視聴量変化に関するラベル定義

コーナー単位の視聴量変動を予測することが目的であるため，各コーナーの視聴量が前のコーナーの視聴量からどう変化したかをもとに増加と減少の 2 ラベルを定義する．このラベルを視聴変化ラベルとする．

増加ラベル
: 前コーナーから視聴量が増加または視聴量の変動が小さく，番組の平均視聴量を超えているコーナー

減少ラベル
: 前コーナーから視聴量が減少または視聴量の変動が小さく，番組の平均視聴量を下回っているコーナー

4　番組コーナーごとの視聴量変化予測

4.1　BERT によるテキストと視聴者属性を統合した分類モデルの構築

単一のモデルで全視聴者属性の視聴変化ラベルを予測する方法と，視聴者属性ごとに視聴変化を予測する方法を比較する．

はじめに単一のモデルでテキストと視聴者属性を用いた分類モデルの定義をする．コーナー内容の要約テキスト（メモ）と視聴者属性を [sep] でつないだ 2 文を入力とし，当該コーナーのその視聴者属性に対する視聴量が増加/減少するかを BERT[6] で予測する（図 2）．視聴者属性はカテゴリであるが，M1 は「20-34 歳男性」を意味しており，これをテキストとして扱う．このモデル構造は BERT に分類層 classifier を加える BertForSequenceClassification と同様のアーキテクチャである．

次に視聴者属性ごとのモデルを説明する．この分類モデルは視聴者属性別に BERT と分類器を用意し (図 3)，同一のコーナー内容の要約テキスト（メモ）を用いて視聴者属性ごとの視聴変化ラベルを予測する．

学習設定について説明する．対象とした特定番組の 3 年間分のデータを使用する．学習データとテス

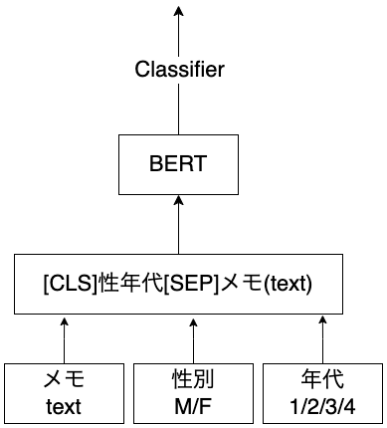

図 2　視聴者属性ごとの視聴量変化ラベル予測モデル: 単一モデル

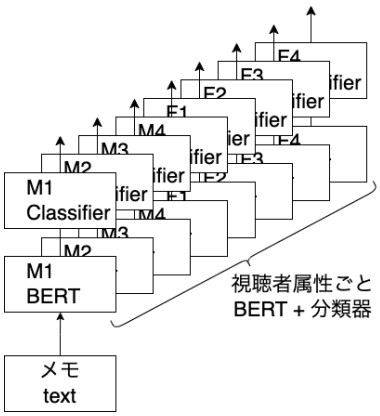

図 3　視聴者属性ごとの視聴量変化ラベル予測モデル: 複数モデル

トデータの分割は 7 対 3 とし，過去の放送回で学習し，未来の放送回でテストする．BERT の事前学習済みモデルは東北大学が公開している事前学習済みモデル [7] を用いる．損失関数は crossentropyloss を使用し，バッチサイズ 8，エポック数 8，最適化関数 AdamW で学習率 1e-5 とした．本研究ではモデルの精度の比較には各モデルそれぞれ 3 回実行した平均で評価する．

視聴者属性別の正解率をモデル別にまとめた（図 4）．視聴者属性別に BERT モデルを用意するより単一のモデルで視聴者属性ごとの視聴変化の予測を行う方が予測精度が高い結果であった．単一のモデルによる正解率は M4,M3,M2 は 75%から 80%程度，F4,F3,F2 は 55%から 60%程度，M1,F1 は 50%前後であった．このような視聴者属性別に正解率に差があった要因は表 2 のように視聴者属性別に各クラスのサンプルサイズに偏りがあることが原因ではないかと考えた．視聴増加のラベル数割合が高い視聴者属性ほど正解率が高かった．

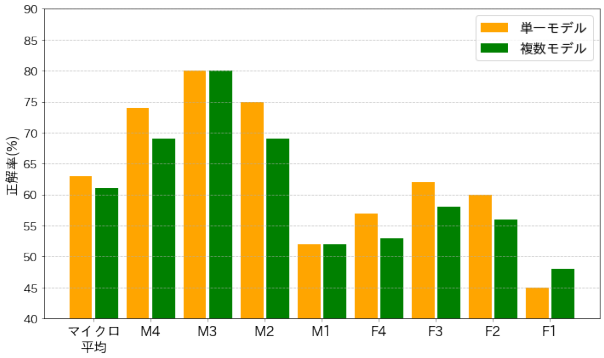

図 4　モデル：入力別の正解率の比較

表 2　視聴者属性別のラベル数割合

	M4	M3	M2	M1	F4	F3	F2	F1
増加	65%	70%	67%	56%	59%	62%	55%	53%
減少	35%	30%	33%	44%	41%	38%	45%	47%

以降の実験ではこの単一モデルをベースモデルとして使用し，精度改善アプローチに取り組む．

4.2　ラベル不均衡解消

本節では前節の結果を踏まえて，ベースモデルに不均衡データの対処をする．本研究のデータでアンダーサンプリングをした場合にテストデータの件数が少なくなるため，オーバーサンプリング手法である SMOTE [8] や Random Over Sampler を実施する．

テキストデータのオーバサンプリング，データ拡張の研究事例を示す．文書内の単語を同義語に置き換える方法や文の意味が変化しない程度に単語の追加や削除などのノイズを加える方法 [9] がある．しかし，本研究でのテキストデータは視聴者属性に紐づいているため本研究でのテキストのデータ拡張に適さないと考えた．また放送内容に関する擬似的なテキストデータを作成することはコストは高い．画像のデータ拡張であればサイズの拡大縮小，回転や位置ずらしなどによる方法があるがテキストではそれが難しい．解決方法としてテキストと視聴者属性を埋め込んだ空間上のベクトルを増やすことを考えた．つまり BERT モデルが出力する pooler-output をデータ拡張する．この方法であれば，既存の機械学習モデルにおけるデータ拡張手法を応用できる．

本研究でテキストデータのデータ拡張と学習を図 5 のように行った．はじめに 4.1 節同様に BERT モデルと分類器の学習を行う．学習後の BERT モデルに学習データ（視聴者属性，メモ）を入力し pooler-output ベクトルを得る．得られた pooler-output ベクトルを用いてオーバーサンプリングする．そしてオーバーサンプリング後のベクトルを用いて分

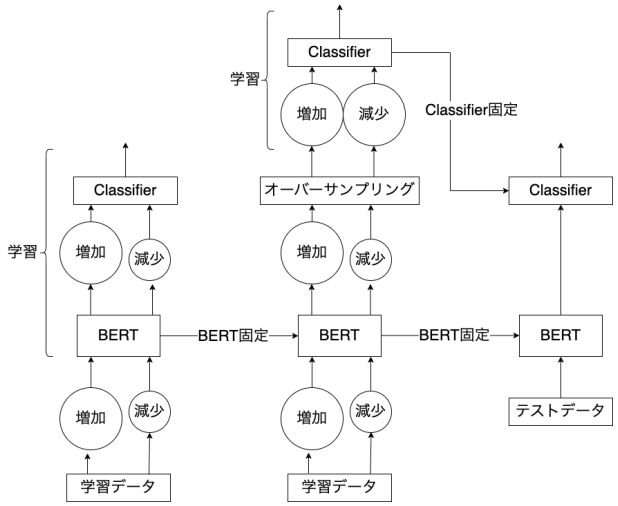

図 5　オーバーサンプリングによるデータ拡張と学習方法

類器のみを再学習する．テストははじめに学習した BERT モデルと再学習した分類器を用いる．

学習の実験設定について説明する．はじめに行う BERT モデルと分類器の学習は 4.1 節と同様の設定である．オーバーサンプリング手法である Random Over Sampler と SMOTE の実装には imbalanced-learn パッケージを用いている [10]．オーバーサンプリング後の分類器の再学習ではエポック数 10 としている．

4.3　BERT に対照学習を入れたコーナーごとの視聴変化ラベル予測

本節ではモデルの学習時に対照学習 [11] を行う．4.2 節のラベル不均衡の解消は本節では適用しておらず，精度改善のための別アプローチとして対照学習を実施している．

対照学習とは距離学習の枠組みの一つであり，ある基準となるデータに対して同一ラベルのデータを近くに，異なるラベルのデータを遠くに埋め込まれるように学習する手法である．対照学習の損失は Triplet Margin Loss を用いる [12]．これは基準となるデータ (anchor) に対する正例 (positive) と負例 (negative) のそれぞれの距離の差を損失関数（式 (1)）とした学習である．本節での a (anchor)，p (positive) と n (negative) は BERT の pooler-output ベクトルである．$d(x, y)$ は x と y のユークリッド距離であり，margin はハイパーパラメータである．

$$L(a, p, n) = \max\{d(a, p) - d(a, n) + \text{margin}, 0\} \quad (1)$$

本節の学習設定について説明する．crossentropy-loss に Triplet Margin Loss を加えたものを損失関数とする．Triplet Margin Loss の margin は 1 とした．そ

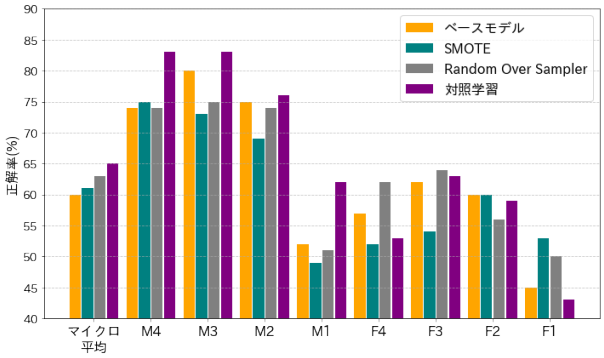

図 6　精度比較

のほかの学習設定は 4.1 節同様の設定である．

4.4　各モデルの結果の比較

ベースモデル，不均衡データの対処と対照学習の正解率を比較する（図 6）．

不均衡データの対処の結果，SMOTE と Random Over Sampler を用いたいずれの場合でも，ラベル不均衡度合いが一番大きかった M3 では正解率がやや低下したが，その他の視聴者属性の正解率は同等もしくは向上する結果となった．視聴者属性別正解率のマイクロ平均ではベースモデルと比較して向上した．視聴者属性ごとにラベルを均衡にした上で分類器を再学習したことで精度の向上を確認することができた．

対照学習を取り入れた場合においても，聴者属性別正解率のマイクロ平均はベースモデルと比較して向上した．特に M3 と M4 の視聴者属性で 80 %を超える結果となった．

ラベル不均衡の解消を行ったうえで対照学習をすることは今後の課題とする．

5　おわりに

本研究では，番組コーナーごとの視聴量変化予測に取り組んだ．4.2 節のオーバーサンプリングと 4.3 節の対照学習を組み合わせた学習方法は今後の取り組むべき課題である．

本研究では 3 年間分のデータを用いてコーナー予測を実施したが，さらにデータ数を拡充したり，ほかの番組においてもコーナー予測をしたりする必要があると考えている．本研究では予測結果に基づき CM チャンスを短くしたり長くしたりした際の影響については考慮していないため，この検証も今後の課題である．

謝辞

本研究は TVS REGZA 株式会社様から視聴データ，株式会社エム・データ様から番組シーンメタを提供していただき研究を行なった．

参考文献

[1] 民放の視聴データに関する取組みについて，（2023-12-15 閲覧）．https://www.soumu.go.jp/main_content/000747672.pdf.

[2] 菊池匡晃, 坪井創吾, 中田康太. 大規模テレビ視聴データによる番組視聴分析. デジタルプラクティス, Vol. 7, No. 4, pp. 352–360, 2016.

[3] 水岡良彰, 中田康太, 折原良平. 大規模テレビ視聴データによる視聴パターン推移の分析. 人工知能学会全国大会論文集 第 32 回 2018, pp. 1P203–1P203.

[4] 東芝テレビ視聴データ分析サービス，（2023-12-15 閲覧）．https://www.regza.com/tvdata.

[5] TV META DATA TV メタデータとは，（2023-12-15 閲覧）．https://mdata.tv/metadata/.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186. Association for Computational Linguistics, June 2019.

[7] 東北大学自然言語処理研究グループ: BERT-base Japanese Whole-Word Masking，（2023-12-15 閲覧）．https://huggingface.co/cl-tohoku/bert-base-japanesewhole-word-masking.

[8] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority oversampling technique. **Journal of artificial intelligence research**, Vol. 16, pp. 321–357, 2002.

[9] Jason Wei and Kai Zou. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 6382–6388, 2019.

[10] Guillaume LemaÃŽtre, Fernando Nogueira, and Christos K Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. **Journal of machine learning research**, Vol. 18, No. 17, pp. 1–5, 2017.

[11] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 6894–6910. Association for Computational Linguistics, November 2021.

[12] TripletMarginLoss, （2023-12-15 閲覧）．https://pytorch.org/docs/stable/generated/torch.nn.TripletMarginLoss.html.