

- 1. はじめに
- 4. 提案手法
- 1. はじめに
- 2. 関連研究?
- 3. BERT と SHAP、クラスタリングについて
- 4. 提案手法
- 5. 数値実験ならびに考察
- 6. おわりに

方向性と進捗

水上 和秀 (Kazuhide Mizukani)
u355020@st.pu-toyama.ac.jp

富山県立大学 工学部 電子情報工学専攻

November 12, 2024

1. はじめに

2/31

研究の背景 (仮)

- SNS や Web サイトの利用者は増加し、個人のインターネット利用率は 80 % を越えており、大勢の人がインターネット上の情報に触れることができるようになった。
- 現代の消費行動の特徴としてインターネット上の口コミや評価を重視し、意思決定や行動に大きな影響を与えている。
- しかしインターネット上の口コミは対面のクチコミよりも利用可能な非言語的手がかりが乏しく、相手に伝わりにくい。

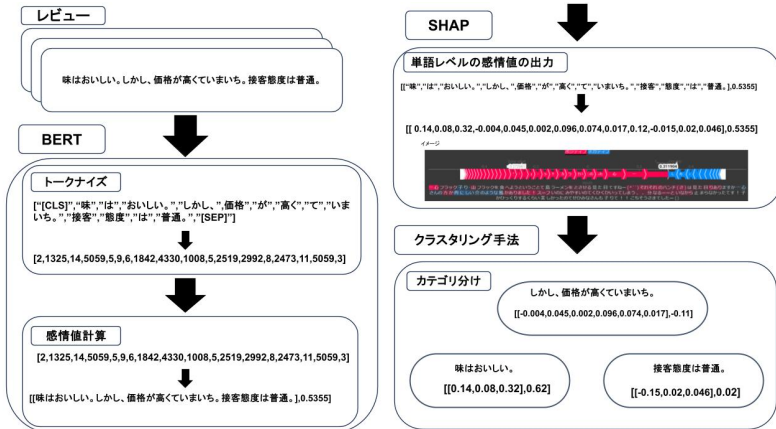
研究の目的 (仮)

- BERT でレビューの感情分析を行い、分析結果に大きく影響を及ぼしている単語を SHAP で可視化を行う。
- トピックごとにクラスタリングを行うことでユーザにとってより解釈可能なレビューの分析を行うシステムを提案する。

実装の流れ

- 1 データセットの取得**
→解きたいタスクの学習用データに使用。
ネガポジ分析の場合、ラベル付けしたネガティブな文章とポジティブな文章を大量に用意する
- 2 分析レビューの取得**
→感情分析を行うためのレビューをスクレイピングする
- 3 事前学習モデルの構築 (BERT)**
→「事前学習」「ファインチューニング」の二つの学習を行うモデルを構築する
- 4 トークナイザーの構築 (BERT)**
→文章を語彙（トークン）に分割したうえで、BERT モデルに入力できる形に変換する処理
- 5 BERT による感情分析**
- 6 SHAP の実装**
→ BERT で分析した結果を可視化 (数値化) する
- 7 レビュー文をクラスタリング**
- 8 出力**
→可視化した結果をわかりやすく表示する

1. はじめに
4. 提案手法
1. はじめに
2. 関連研究?
3. BERT と SHAP、クラスタリングについて
4. 提案手法
5. 数値実験ならびに考察
6. おわりに



できてること

- レビューサイトのレビューをスクレイピング
- データセットの収集
- 自分で BERT モデルの構築
- SHAP の実装

問題点

- 感情分析の精度がまだよくない
- レビュー文のクラスタリングについて

まだ

- システム化する (サーバー上でできるようにする)
- ほかの機械学習の手法と比較
- わかりやすく出力する

1. はじめに
4. 提案手法
1. はじめに
2. 関連研究?
3. BERT と SHAP、クラスタリングについて
4. 提案手法
5. 数値実験ならびに考察
6. おわりに

1. はじめに
4. 提案手法
1. はじめに
2. 関連研究?
3. BERT と SHAP、クラスタリングについて
4. 提案手法
5. 数値実験ならびに考察
6. おわりに

進捗

- レビューの整形
- bert のパラメータの調整
- レビュー文クラスタリングの検討

[illegible]

BERT のパラメータについて

以下は BERT に用いるパラメータ。どのパラメータが分析精度に影響しているか検討中。

- エポック数: トレーニングデータを学習させる回数
- バッチサイズ: 一度にモデルに与えるデータの数
- ウォームアップステップ: 学習率を徐々に増加させるためのステップ数
- 重みの減衰率: パラメータ更新の際に重みの値を少しずつ減らす手法

```
1 # モデルの初期化
model = BertForSequenceClassification.from_pretrained('bert-base-uncased', num_labels=2)

# トレーニング引数の設定
training_args = TrainingArguments(
    output_dir='./results',
    num_train_epochs=3, #エポック数
    per_device_train_batch_size=16, #とレーニングのバッチサイズ
    per_device_eval_batch_size=16, #評価のバッチサイズ
    warmup_steps=500, #ウォーミングステップ
    weight_decay=0.01, #重みの減衰率
    logging_dir='./logs',
    logging_steps=10,
)
```

- 1. はじめに
- 4. 提案手法
- 1. はじめに
- 2. 関連研究?
- 3. BERT と SHAP、クラスタリングについて
- 4. 提案手法
- 5. 数値実験ならびに考察
- 6. おわりに

クラスタリングについて

文章をカテゴリごとに分類するために、クラスタリング手法を用いる。ラベル付き教師用データがないので、教師なし学習によるクラスタリング手法を用いる。

kmeans 法と LDA 法

kmeans 法はデータを数値化し、互いに近いデータを同じクラスに分類するクラスタリング手法。

LDA は、文書内の「トピック」を潜在的に見つけ出し、そのトピックに基づいて文書を分類する手法。LDA では、各文書は複数のトピックの混合であると仮定し、トピックごとの単語の出現確率分布を計算する。

実装してみた

レビューを文節ごとに区切り、kmeans 法で実際にクラスタ分けを行った。
(味、価格、接客態度、その他) に分類されることを目的に、クラスタ数は 4 つに設定を行った。

[illegible]

1. はじめに

4. 提案手法

1. はじめに

2. 関連研究?

3. BERT と
SHAP、クラスタ
リングについて

4. 提案手法

5. 数値実験ならび
に考察

6. おわりに

考察

- kmeans 法は似たような文章を自動でクラスタリングする手法だが、思ったようなクラスタ分けができていなかった→レビュー文を適正な方法で数値化する必要がある
- また、クラスタ数を適切に決める必要がある

1. はじめに

4. 提案手法

1. はじめに

2. 関連研究?

3. BERT と
SHAP、クラスタ
リングについて

4. 提案手法

5. 数値実験ならび
に考察

6. おわりに

付け加えるとしたら?

分析するレビューにさくらレビューが紛れている可能性があるので、さくらレビューの判断 (清水さん)

1. はじめに

4. 提案手法

1. はじめに

2. 関連研究?

3. BERT と
SHAP、クラスタ
リングについて

4. 提案手法

5. 数値実験ならび
に考察

6. おわりに

やったこと

- レビューの整型を行った
- BERT のパラメータを調整した
- レビュー文クラスタリングを行った

これからやること

- BERT のモデルの調整
- それぞれの精度を上げる
- システム化

1. はじめに
4. 提案手法
1. はじめに
2. 関連研究?
3. BERT と SHAP、クラスタリングについて
4. 提案手法
5. 数値実験ならびに考察
6. おわりに

以下、前回の資料

1. はじめに

15/31

研究の背景 (仮)

- SNS や Web サイトの利用者は増加し、個人のインターネット利用率は 80 % を越えており、大勢の人がインターネット上の情報に触れることができるようになった。
- 現代の消費行動の特徴としてインターネット上の口コミや評価を重視し、意思決定や行動に大きな影響を与えている。
- しかしインターネット上の口コミは対面のクチコミよりも利用可能な非言語的手がかりが乏しく、相手に伝わりにくい。

研究の目的 (仮)

- BERT でレビューの感情分析を行い、分析結果に大きく影響を及ぼしている単語を SHAP で可視化を行う。
- トピックごとにクラスタリングを行うことでユーザにとってより解釈可能なレビューの分析を行うシステムを提案する。

2. 関連研究?

16/31

- 1. はじめに
- 4. 提案手法
- 1. はじめに
- 2. 関連研究?
- 3. BERT と SHAP、クラスタリングについて
- 4. 提案手法
- 5. 数値実験ならびに考察
- 6. おわりに

2.1

2.2

2.3

3.1 BERT について?

17/31

- 1. はじめに
- 4. 提案手法
- 1. はじめに
- 2. 関連研究?
- 3. BERT と SHAP、クラスターリングについて
- 4. 提案手法
- 5. 数値実験ならびに考察
- 6. おわりに

BERT とは

Bidirectional Encoder Representations from Transformers (BERT) は、Devlin らが提案した自然言語処理モデルであり、双方向 Transformer というニューラルネットワークを利用している。BERT は大量の教師なしデータによる事前学習を行ったモデルを、比較的少量の教師ありデータでファインチューニングして個々のタスクに合わせて学習する。これにより事前学習済みモデルをそのまま適用するよりも性能の向上が期待できる。

また、BERT の特徴として、文章を文頭と文末の双方向から学習することで、文脈を考慮できることが挙げられる。BERT ではこの双方向からの学習を、「事前学習」の中の「MLM」「NSP」の2つのタスクを行うことで可能としている

3.1 BERT について?

18/31

- ある文章において一部のトークンを特殊トークンである [MASK] に置き換えて、その [MASK] に入るトークンを予測する言語モデルのこと。
→ 単語に対応する文脈情報を獲得できる
- 「意味的につながりがある 2 文」、または「意味的につながりのない 2 文」を入力し、2 文の関係性を考慮することで「入力された 2 文に意味的につながりがあるかどうか」を予測する
→ 単語の関係性だけでなく文章の関係性の情報も獲得できる

① 一部分が [MASK] で置き換えられた文を入力
(置き換えられる前の単語は「野球」)

私の好きなスポーツは [MASK] で、いつか本場のアメリカで観戦してみたい。

入力



② 前後の単語の並びから [MASK] で置き換えられた単語を予測

出力

野球

① 意味的につながりがある 2 文や
つながりがない 2 文を入力

1文目: 私は [MASK] とアメリカに行った。
2文目: そこで [MASK] に英語で話しかけられた。

入力



② 2 文の関係性から意味的につながりがあるかどうか予測

出力 → つながりがある

1文目: 私は [MASK] とアメリカに行った。
2文目: イルカは [MASK] の哺乳類だ。

入力

出力 → つながりがない

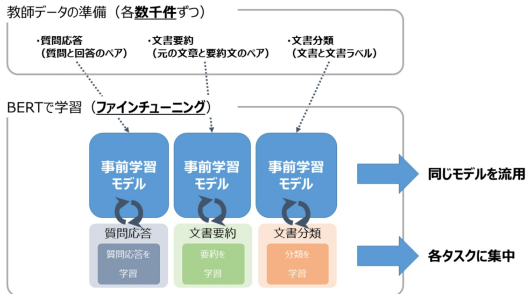
3.1 BERT について?

19/31

ファインチューニング

BERT の学習済みモデルは、そのまま使うことは珍しく、一般に、解きたいタスクに応じて特化するようにする。

ファインチューニングを行うときにはモデルの初期値として、事前学習で得られたパラメータを用い、新たに加えられた分類器のパラメータにはランダムな値を与える。そして、ラベル付きデータを用いて BERT と分類器の両方のパラメータを学習する → ファインチューニングの際事前学習で得られたパラメータを初期値として用いることで比較的少数の学習データでも高い性能のモデルを得ることができる



1. はじめに
4. 提案手法
1. はじめに
2. 関連研究?
3. BERT と SHAP、クラスタリングについて
4. 提案手法
5. 数値実験ならびに考察
6. おわりに

3.2 SHAP について?

SAHP とは

機械学習で導出した予測値に対して各特徴量がどのくらい寄与しているかを算出する手法で、シャープレイ値の考え方に基づいている

シャープレイ値とは

協力ゲーム理論において複数のプレイヤーの協力によって得られた利得を各プレイヤーの貢献度に応じて構成に分配するための手段の一つ

- 3 人のプレイヤー (1.2.3) が協力してゲームに挑戦し、利得として、以下の賞金
が得られるとする
- このときの 1.2.3 にそれぞれどのようにお金を分配するか。

表1 協力ゲームの例

参加プレイヤー	賞金/万円
1	4
2	6
3	10
1, 2	16
1, 3	22
2, 3	30
1, 2, 3	60

シャープレイ値とは

21/31

- このとき各プレイヤーの限界貢献度を導入する。限界貢献度とは、プレイヤー i が参加したときの利得の増加分である。
- 例えば、プレイヤーの参加順「 $1 \rightarrow 2 \rightarrow 3$ 」のときのプレイヤー 3 の限界貢献度は、 $v(1, 2, 3) - v(1, 2) = 60 - 16 = 44$ のように計算できる。
- 各プレイヤーのシャープレイ値は以下のようになる

プレイヤー 1: $(4 + 4 + 10 + 30 + 12 + 30)/6 = 15$

プレイヤー 2: $(12 + 38 + 6 + 6 + 38 + 20)/6 = 20$

プレイヤー 3: $(44 + 18 + 44 + 24 + 10 + 10)/6 = 25$

表2 限界貢献度

プレイヤーの参加順	各プレイヤーの限界貢献度		
	1	2	3
$1 \rightarrow 2 \rightarrow 3$	4	12	44
$1 \rightarrow 3 \rightarrow 2$	4	38	18
$2 \rightarrow 1 \rightarrow 3$	10	6	44
$2 \rightarrow 3 \rightarrow 1$	30	6	24
$3 \rightarrow 1 \rightarrow 2$	12	38	10
$3 \rightarrow 2 \rightarrow 1$	30	20	10

- はじめに
- 提案手法
- はじめに
- 関連研究?
- BERT と SHAP、クラスタリングについて
- 提案手法
- 数値実験ならびに考察
- おわりに

1. はじめに
4. 提案手法
1. はじめに
2. 関連研究?
3. BERT と SHAP、クラスタリングについて
4. 提案手法
5. 数値実験ならびに考察
6. おわりに

シャープレイ値の定式化

- 一般的には、プレイヤー i のシャープレイ値は次式によって定式化される。ただし、 s は連携 S に含まれるプレイヤー数である。

$$\phi_i = \sum_{S: i \in S \subset N} \frac{(s-1)!(n-s)!}{n!} \{v(S) - v(S - \{i\})\}$$

シャープレイ値と SHAP について

協力ゲーム理論のシャープレイ値の概念を応用して、特徴量の貢献度を計算

協力ゲーム (Shapley値)

貢献が違う複数人で仕事をしたときに
各人の仕事に対する貢献度を考慮して
公平に報酬を分配する方法



SHAP

貢献が違う複数の特徴量で予測を
したときに、各特徴量の予測に対
する貢献度を考慮して、公平に
予測値への貢献を分配する方法



SHAP の定式化

解釈したい予測モデルを f 、バイナリ変数 (0 か 1 の変数) を z 、各特徴量に対する貢献度を ϕ_i とすると以下のようにあらわす

$$g(z) = \sum_{i=1}^p \phi_i z_i$$

$$\phi_i(f, x) = \sum_{z \subseteq x} \frac{|z|!(p - |z| - 1)!}{p!} [f(z) - f(z \setminus i)]$$

1. はじめに
4. 提案手法
1. はじめに
2. 関連研究?
3. BERT と SHAP、クラスタリングについて
4. 提案手法
5. 数値実験ならびに考察
6. おわりに

1. はじめに
4. 提案手法
1. はじめに
2. 関連研究?
3. BERT と SHAP、クラスタリングについて
4. 提案手法
5. 数値実験ならびに考察
6. おわりに

文章のカテゴリ分類について

- レビューの文章を文節ごとに区切り、カテゴリ (味に関する文、価格に対する文、接客態度に関する文、など) ごとにクラスタリング分けを行う。

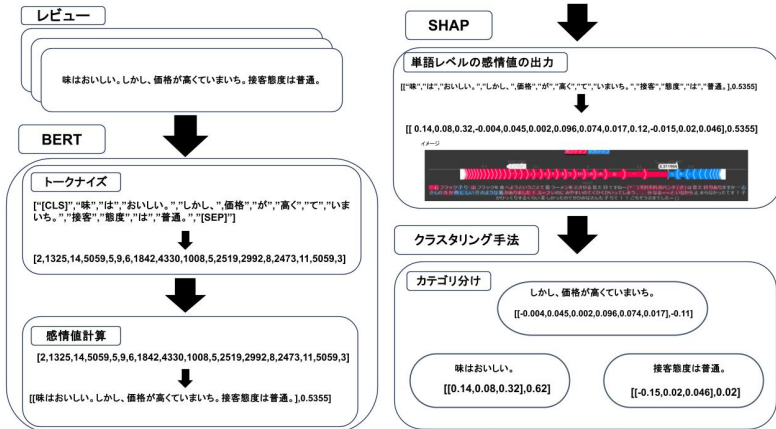
クラスタリング手法

クラスタリング手法で定番の k-means 法か LDA を検討中

実装の流れ

- 1** データセットの取得
→解きたいタスクの学習用データに使用。
ネガポジ分析の場合、ラベル付けしたネガティブな文章とポジティブな文章を大量に用意する
- 2** 分析レビューの取得
→感情分析を行うためのレビューをスクレイピングする
- 3** 事前学習モデルの構築 (BERT)
→「事前学習」「ファインチューニング」の二つの学習を行うモデルを構築する
- 4** トークナイザーの構築 (BERT)
→文章を語彙 (トークン) に分割したうえで、BERT モデルに入力できる形に変換する処理
- 5** BERT による感情分析
- 6** SHAP の実装
→ BERT で分析した結果を可視化する
- 7** レビュー文をクラスタリング
- 8** 出力
→可視化した結果をわかりやすく表示する

1. はじめに
4. 提案手法
1. はじめに
2. 関連研究?
3. BERT と SHAP、クラスタリングについて
4. 提案手法
5. 数値実験ならびに考察
6. おわりに



進捗 1(データセットの取得)

- 1 TIS が公開している感情分析を行うためのデータセット「chABSA-dataset」を取得できるようにした。
- 2 「chABSA-dataset」は上場企業の有価証券報告書(2016年度)をベースに作成されたデータセットで、各文に対してネガティブ・ポジティブのラベル付けがされている
- 3 このデータセットを BERT に読み込ませ、モデル構築を行う

text	labels
売上収益は、抗がん剤「レンビマ」(欧州における腎細胞がんに係る製品名「Kisplyx」)および抗てん	1
セグメント別には、日本医薬品事業およびアジア医薬品事業が増収となりました	1
また、すべての海外セグメントにおいて現地通貨ベースで成長を果たしました	1
グローバルブランド4品目合計では、為替の影響を受ける中で、前期から14.5%増の728億22百万円と	1
営業利益は、オペレーションの高質化・効率化に加え、EAファーマ株式取得に伴う一時収益(割安購	1
親会社の所有者に帰属する当期利益は、393億58百万円(前期比28.4%減)となりました	0
基本的1株当たり当期利益は、137円63銭(前期より54円61銭減)となりました	0
当期利益にその他の包括利益を加減した当期包括利益は、前期に円高の進行によって為替換算差額が	1
売上収益の内訳は、医療用医薬品が2,439億99百万円(同4.3%増)、ジェネリック医薬品が280億27百万	1
品目別売上収益については、ニューロロジー領域で、不眠症治療剤「ルネスタ」が80億12百万円(前期	1
ファイザー社と共同販促を展開している疼痛治療剤「リリカ」の共同販促収入は242億68百万円(同1.8	0
オンコロジー領域では、「ハラヴェン」が77億63百万円(同14.2%増)、「レンビマ」が27億15百万円	1
さらに、ヒト型抗ヒトTNF α モノクローナル抗体「ヒュミラ」は376億62百万円(同15.4%増)と順調に	1
売上収益は、1,172億17百万円(前期比4.1%減、現地通貨ベースでは6.3%増)となりました	0
セグメント利益は、効率的なマーケティング活動による販売管理費の減少および前期に発生した米国	1
品目別売上収益については、ニューロロジー領域で、抗てんかん剤「Banzel」が138億48百万円(前期	1
「Belviqu」は37億14百万円(同16.0%減)でした	0
オンコロジー領域では、制吐剤「Aloxi」が480億82百万円(同12.1%減)、「ハラヴェン」が166億19百	0
売上収益は492億74百万円(前期と同等水準、現地通貨ベースでは17.0%増)、セグメント利益は136億96	1

1. はじめに
4. 提案手法
1. はじめに
2. 関連研究?
3. BERT と SHAP、クラスタリングについて
4. 提案手法
5. 数値実験ならびに考察
6. おわりに

進捗 2(分析レビューの取得)

1. はじめに
4. 提案手法
1. はじめに
2. 関連研究?
3. BERT と SHAP、クラスタリングについて
4. 提案手法
5. 数値実験ならびに考察
6. おわりに

- 1 食べログのラーメンのレビューをスクレイピングし、利用できるようにした。
- 2 1つの店に対して上位 20 件分を取得し、感情分析をして、その店の分析結果を表示する

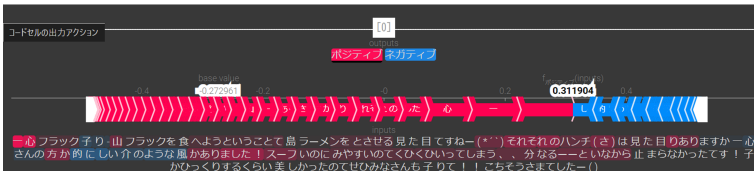
Source Nani Review
貴族車 水見屋煮子ラーメン 1000円今日は、富山県を代表する富山の人気を誇るラーメン屋である貴族さんにお邪魔しました！初見で読むことが不可能なお店の名前にもじんなに旅行中、通りかかった水見にて美味しさを伝えたところ、断った点の客が店内に遭遇、すぐに向かいました。平日13:20:15の列に接続、すでに由緒長いやぶさ貴族の朝に伺いました。6時半到着で7番目まで入行。朝は涼しいので待ち時間苦しくなりました。1回目でお店。7時には20人以上の待ち。スプーンが大量実用かつ、自覚めたやぶさ貴族 富山、富山旅行の初日は水見！湯通と丼物と思いつき、折角のため10名ほど見てみませんか？という事で水見駅からタクシーで店へ。11時45分到着する58人(赤ちゃんと合)待ち 今日のお店は富山県水見町の貴族(としん)さんにお伺いします。営業時間：11:00-15:00 / 23:00-23:00 定休日：月曜日 座席：カウンター10席 / テーブル席 駐車場 貴族 〇に入った点・旨みだけを抽出した透明感のある煮汁スープ〇追加情報：御一考があった専用メニューのように：難車乗る〇ルール、列は道路を挟んだ向かい側の貴族 〇に入った点・旨みだけを抽出した透明感のある煮汁スープ・様々な具材の彩りある旨味をまよくまとめた白濁スープ〇追加情報：1時間待ち待たせ方をしよう。貴族 〇に入った点・旨みだけを抽出した透明感のある煮汁スープ●注意：1時間は待つ覚悟をしましょう。〇：駐車場の〇ルール、列は道路を挟んだ向かい側の貴族 貴族に訪ねました。今回朝ラーメンしているのとこの朝に富山風ラーメンを注文したスプーンは嫌いではないかともしりより美味しくてつては糖は糖味を使用して 貴族風に訪問白濁油外売りに付く状態です……………湯通しに介介が効いて美味しく使います。また、白濁油にコクがあり良いです。福中は味噌味噌使用されています。チャ 水見屋煮子ラーメン1000貴族富山県水見区に移住して「貴族」に念願の初訪帳。店の前に二人、向かいに一人居ましたが店業案内係実質無し。営業家オスメスの 貴族 見え3.4スナッ対応3.8券面4.0レベニューあがりつもりが更新されていなかったようであり適当営業家はゆさぶとへの噂においては最近始まった朝ラーカーを試そうの 貴族 水見屋煮子ラーメン1000★★★★★☆☆☆女性のみを取りတဲ့富山旅行。富山県No.1の評価で、水々BMW来たにちなら。平日12時半の時点で客と25人の行列が。1時間前 最高に美味かったですごちようございましたBuono!-私はお店を評価できるような者ではございませんのでジャッジ!人数を付けての併席は控えてさせていただきます。 貴族 最初に美味かったですごちようございましたBuono!-私はお店を評価できるような者ではございませんのでジャッジ!人数を付けての併席は控えてさせていただきます。 貴族 土日は朝ラーメンをやっていると聞きました。こちらう百名店「貴族(としん)」さんです。開店以前前の時から既に4人で着目しています。ちょっと物を置いて 貴族 さっさと来たにきたからと思い、貴族さんへ質問しました。富山駅から高岡駅乗り換えで約60分、水見駅に約10分の距離です。お店には開店2時間前の 貴族 以前から行きたいと思って、貴族(としん)さん。富山のラーメン名店の本店です。水見屋煮子に近くて道幅415号路一本となっており、人気の多い。牛山 富山車「うまい」というのが、その日富山ラーメンを始めたのがここだった。顔の赤さでなく、口の中が辛く辛いという感じがたまにたまに、開店店

進捗 3(BERT モデルの構築、SHAP の実装)

- 1 自分でモデルを作成し、レビュー文に適応させ、感情値を出力できるようにした。
- 2 BERT で出力した結果対して、SHAP で可視化できるようにした。

※ テストデータ

```
test_texts = ["***【一心ブラック煮玉子入り ¥980-】富山ブラックを食べようということで訪問徳島ラーメンを彷彿とさせる見た目ですね(・_・)味それぞれ醤油のパンチ(濃さ)は見た目通りありますが一心さんの方が全体的に優しい魚介のような風味がありました！スープ濃いのに飲みやすいのでくびくびいってしまう、塩分過多なるーと思いながら止まらなかったです！煮玉子がびっくりするくらい美味しかったのでぜひみなさんも煮玉子入りで！！ごちそうさまでしたー(・_・)"]
test_encodings = tokenizer(test_texts, truncation=True, padding=True, max_length=128, return_tensors='pt')
```



- 1 出力結果の精度がまだよくないので、モデルの設定を変えて精度を上げる
- 2 文章をカテゴリごとに分類できていないため、レビュー文をカテゴリごとに分類する手法を検討

- 1. はじめに
- 4. 提案手法
- 1. はじめに
- 2. 関連研究?
- 3. BERT と SHAP、クラスタリングについて
- 4. 提案手法
- 5. 数値実験ならびに考察
- 6. おわりに

有効性の検証

- 1 ほかの感情分析の手法も実装し、どれだけ優れているか比較する
- 2 実際にシステムを利用してもらい、どれだけつかいやすかったかアンケートを取る
- 3 ほかに有効性を示す手法があれば検討

6. おわりに

31/31

1. はじめに
4. 提案手法
1. はじめに
2. 関連研究?
3. BERT と SHAP、クラスタリングについて
4. 提案手法
5. 数値実験ならびに考察
6. おわりに

やったこと

- あ

これからやること

- い