

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

September 20, 2024

研究の背景 (仮)

- レビューの文章のネガポジ分析
→ BERT と SHAP を用いてレビュー文の単語レベルでの感情分析を行い、その結果を可視化するシステムを提案

研究の目的 (仮)

- 1 レビュー文のスクレイピング
- 2 BERT による文章の感情値を分析
- 3 SHAP による分析結果の解釈
- 4 結果の出力

BERT とは

BERT は、Google が開発した自然言語処理のための深層学習モデル。文の前後を考慮した文脈理解ができる。これにより、高度な自然言語処理タスクを行うことが可能になる。

BERT とは

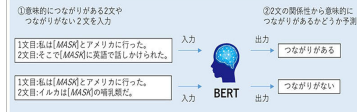
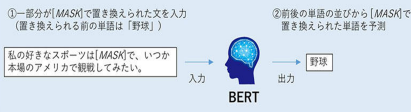
3/15

学習方法

BERT の学習方法は「事前学習」「ファインチューニング」の 2 段階がある。事前学習は、ラベル無しデータを用いて、複数のタスクで事前学習を行うことであり、ファインチューニングは事前学習の重みを初期値として、ラベルありデータでファインチューニング（微調整）を行う。

BERT では MLM と NSP の二つの事前学習を行う

- 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



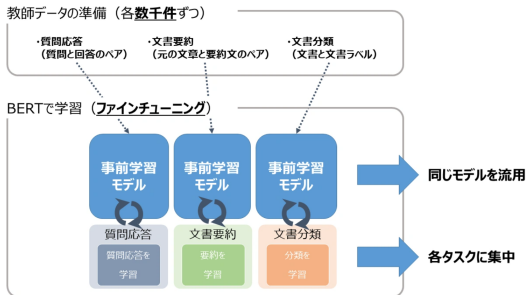
BERT とは

4/15

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



SHAP とは

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

6/15

- 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

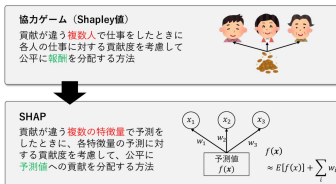
シャープレイ値の定式化

- 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



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

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

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

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 SHAP の実装
→ BERT で分析した結果を可視化する
- 6 出力
→可視化した結果をわかりやすく表示する

実装の流れ (イメージ)

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
「Belvq」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

進捗 2(分析レビューの取得)

- 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

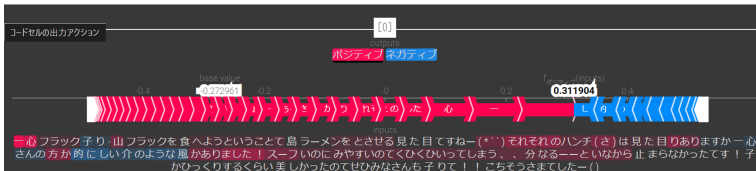
Store Name Review	
貴賓席	水見座敷千ラーメン1000円今日は、富山県を代表する富山1の人気を誇るラーメン屋である貴賓席さんにお邪魔しました！ 初見！読むことが不可能なお店の名前はとんじろ
貴賓席	旅行中、通りかかった水見に不味いお店を探していたところ、断り1点数のこちらに遭遇、すぐに向きました。平日13:20に15人の列に接続、すぐに店員さんがやってきました。朝の朝に相いました。6時半到着で7時とあつた。朝は12時まで待ちを苦みかまふました。12時で入店。7時に20人以上の列が並ぶ。スープが大量美味し、自宅までた
貴賓席	日曜、富山市の初日は水見1 湯地と物と思いつ、折角のため100名店員を苦みかまふた？という事で水見駅前のタクシーまで。11時48分に到着すると54人(赤いふく)待ち
貴賓席	今日のお昼は富山県水見市の貴賓席(とんじろ)さんにお伺いします。営業時間：11:00-15:00 / 17:00-23:00 定休日：月曜日 夜店：カウンター10席 / テーブル8席 座席
貴賓席	〇氣に入った点・旨みだけを抽出した透明感ある煮干しスープ〇追加情報・朝方が始まるとが専用メニューのよう〇駐車場有〇丸・列は待つ倍を挟んだ側から割
貴賓席	〇氣に入った点・旨みだけを抽出した透明感ある煮干しスープ・様々な具材の煮込みが旨味をもつとめとの白濁スープ〇追加情報・1時間待ち待倍をしよう！
貴賓席	〇氣に入った点・旨みだけを抽出した透明感ある煮干しスープ〇注意・1時間待ち待倍をしよう！
貴賓席	貴賓席に訪問しました。今回お昼に貴賓席でいること〇朝に訪問風ラーメンを注文したスープは美味しいとありしり美味しとては難は難を使用して
貴賓席	貴賓席に訪問白濁油外売り状態です……………湯地汁を介が効いてあり美味しです。また、白濁油にクが好いります。福中経理使用感強しは営業です。チャ
貴賓席	水見座敷千ラーメン1000貴賓席〇水見水見に移動して富山1に念願の初訪問。店の前に二人、向かい一人並ぶましたが直ぐ案内され美味し無し。営業しますの
貴賓席	※3.4タップ付3.8タップ4.0レビューあげたつもり更新されていなかったやうで直し湯満営業はめつとん並ぶの噂を聞いていたので最速はまた朝方から拭きと
貴賓席	水見座敷千ラーメン1000★★★★★☆☆☆☆みを取りつて富山1訪。富山県No.1の評価で、永らくBMにてたち。平日12時半の時間帯に着くと25人の行列が。1時間
貴賓席	最高に美味かつたです！うまかつたBuono!私はお店を評価できるような者ではないのででジャッジし点数を付けての評価は投えてさせていただきます。
貴賓席	最高に美味かつたです！うまかつたBuono!私はお店を評価できるような者ではないのででジャッジし点数を付けての評価は投えてさせていただきます。
貴賓席	土日は朝方ラーメンをやてつと事て来ました。こちらラーメン百名店「貴賓席(とんじろ)」さんです。開店1時間前の6時30分頃に着いて一目でました。ちょっと物を置いて
貴賓席	せっかく富山に来たからと…と思い、貴賓席さんへ訪問しました。富山駅から高岡駅乗り換え約60分、水見駅に到着。駅から約10分の距離です。お店は開店2時間前の
貴賓席	以前から行きたーと思ってた、貴賓席(とんじろ)さん。富山のラーメン名店の本店です。水見湯地にはまだててて415号線を入ったところから、人気のあつ。牛
貴賓席	富山県「ラーメン人気No.1」のトピックが今日12時朝方までを始めたので行ってみたい。朝のてそんてに並ばなくてと予想していたがこんなに多く、開店

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 = ["**【一山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ー(*´`)時それぞれ醤油のパンチ(濃さ)は見た目通りありますが一山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

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

やったこと

- レビューサイトのレビューをスクレイピングした
- 自分で BERT モデルを構築した
- BERT を用いて文章の感情 (ポジティブ、ネガティブ) を分類した
- SHAP を用いて、文章のどの部分が感情に強く影響しているか可視化した

これからやること

- BERT のモデルの調整
- レビュー文をカテゴリ (味に対する文、価格に対する文) ごとに分類分けをする
- SHAP の出力結果をより分かりやすく表示できるようにする