

# 都市を横断した市民意見抽出に関する 課題と手法についての検証

石田 哲也<sup>1</sup> 関 洋平<sup>2</sup> 櫻 惇志<sup>3</sup> 柏野 和佳子<sup>4</sup> 神門 典子<sup>5,6</sup>

**概要:** 行政の政策や接客業のサービスの質を向上させるためには、市民によるフィードバックの収集/分析が重要となる。また、都市によって政策やサービスは異なり、市民の抱える意見も異なるため、都市にとらわれない市民意見の分析を行う必要がある。本研究では、アプレイザル理論に基づく意見タイプを含む、複数の属性を用いた市民意見抽出手法を利用して、都市を横断して市民意見を抽出する手法について検証する。実験では、横浜市民と札幌市民のつぶやきを対象として、意見タイプおよび関連する属性を推定する際に、マルチタスク学習によってファインチューニングを行った T5 モデルが有効であることを示した。また、特定の都市のつぶやきでマルチタスク学習によってファインチューニングしたモデルを、評価対象の都市の比較的少量のつぶやきを用いて再度マルチタスク学習によってファインチューニングする手法の有効性を確認した。さらに、評価対象都市のつぶやきのアノテーションは、特定の都市のモデルによる予測の確信度が高いものを選定してから行うことが有効であることを示した。

## 1. はじめに

自治体による政策の改善には、都市で暮らす市民の意見を収集し、反映させることが重要となる。また、接客業のサービスの質を向上させるためにも、店員の接客に対する批評や、提供している商品の評価等の意見を反映させることが重要となる。これらの背景から、著者らの先行研究 [1] では、特定の都市の市民による Twitter<sup>\*1</sup> のつぶやきから、分析対象の市民意見を自動抽出するためのフレームワークを提案した。Twitter では多くのユーザが日頃感じたことを気軽に述べており、市民が生活している都市の自治体による政策や、日頃利用している接客業のサービスに関する多様な市民意見を収集することができる。

著者らの先行研究では、日本で最も多くの市民が暮らす市である横浜市に着目することで、新型コロナウイルス感染症の流行によって休園や登園自粛で大きな問題となった「保育園」に関する市民意見と、同じく新型コロナウイルス感染症の流行によって急激に利用者が増加した「飲食店のテイクアウトサービス」に関する市民意見抽出における手

法の有効性を検証した。著者らの先行研究におけるフレームワークでは、はじめに人手で作成した市民意見分析コーパスを訓練データとして、BERT[2] を用いたマルチタスク学習モデルをファインチューニングすることで、つぶやきの意見タイプや極性等の複数の属性を推定する。そして、これらの複数の属性を組合せて指定することで、条件を満たすつぶやきのみを抽出する。

しかし、都市によって自治体の政策やその都市の店舗の接客業のサービスは異なるため、新しい都市を対象として市民意見を抽出するには、その都市のつぶやきを対象とした人手によるアノテーションが必要となる。一方で、すべての都市を対象として訓練データを作成した場合には、非常にコストが大きくなる。そこで本研究では、訓練データを構築済みの都市（以降、ソース都市と呼ぶ）のデータを活用して評価対象の都市（以降、ターゲット都市と呼ぶ）の市民意見を抽出する手法を提案する。また、新たに政令指定都市の 1 つである札幌市に暮らす市民のつぶやきを追加することで、市民意見分析コーパスを複数の都市に拡張し、作成した複数の都市のつぶやきからなる市民意見分析コーパスを用いて、手法の有効性と課題について検証する。

本研究の貢献は以下の通りである。

- (1) 都市を横断した市民意見抽出手法の有効性の検証のため、横浜市民と札幌市民のつぶやきからなる市民意見分析コーパスを作成した。
- (2) ソース都市のデータとターゲット都市の比較的少量の

<sup>1</sup> 筑波大学大学院人間総合科学学術院人間総合科学研究群情報学学位プログラム

<sup>2</sup> 筑波大学図書館情報メディア系

<sup>3</sup> 一橋大学ソーシャル・データサイエンス教育研究推進センター

<sup>4</sup> 国立国語研究所

<sup>5</sup> 国立情報学研究所

<sup>6</sup> 総合研究大学院大学

\*1 <https://twitter.com>

データを用いて2段階のファインチューニングを行う手法を提案し、各属性の推定における有効性を検証した。

- (3) ターゲット都市のデータのアノテーションは、先にソース都市のデータでファインチューニングしたモデルによる予測の確信度が高いものを選定した後に行うことが有効であることを示した。

## 2. 関連研究

### 2.1 アプレイザル理論

アプレイザル理論 [3] は、選択体系機能言語学 (Systemic functional linguistics) [4] の立場から提案された、意見を体系化する理論であり、アプレイザル理論を用いることで、意見の対象に着目した分析を行うことができる。第二著者の先行研究 [5] では、アプレイザル理論に基づき、コミュニティ QA 文書を用いた意見分析コーパスを手で作成、分析することで、人物・組織に対する批評意見が政治カテゴリの文書に多く含まれることを示した。また、Mora [6] は、アプレイザル理論に基づき、オンラインストアのレビューを分析し、事物に対する評価が多く含まれることを示した。このように、意見の対象は分析対象の分野によって異なるため、対象に着目した意見分析を行う必要がある。本研究では、アプレイザル理論に基づく意見タイプを定義し、複数の意見属性の推定をマルチタスク学習として定式化する。

### 2.2 マルチタスク学習と T5

マルチタスク学習 [7] とは、共通のモデルを用いて関連する複数のタスクを同時に学習することで、各タスクを独立して学習するよりも高い精度を実現できる手法である。マルチタスク学習では、異なるタスクの学習によって得られた情報が帰納バイアス (inductive bias) [8] としての役割を果たすことで、モデルの一般性を向上し、未知のデータに対する予測性能を高めることができる。著者らの先行研究では、BERT を用いたマルチタスク学習手法である、Multi-Task Deep Neural Networks [9] を参考にマルチタスク学習モデルを構築したが、本研究では、BERT と比較して多くのタスクで高い性能を示す、T5 モデル [10] を用いてマルチタスク学習モデルを構築する。T5 は分類タスクや回帰タスク、翻訳タスクなどの全ての自然言語処理タスクの入出力を全てテキストとして統一された形式で扱うことで、単一のモデル構造で多くのタスクにファインチューニング可能であり、マルチタスク学習モデルの構築に適した言語モデルである。

## 3. 提案手法

本研究における市民意見抽出では、はじめに一部のつぶやきに 3.1 節で述べる属性を付与した市民意見分析コーパ

スを手で作成する。続いて、作成した市民意見分析コーパスを教師データとして、各属性の分類モデルを訓練する。そして、訓練したモデルに未知のつぶやきを入力し、属性を推定することで、つぶやきに複数の属性を自動で付与する。これらの属性を任意の数だけ指定することで、未知のつぶやきから指定された条件を全て満たす市民意見のみを抽出することができる。そこで、3.1 節でつぶやきに付与する各属性を定義し、3.2 節で市民意見抽出のためのフレームワークについて説明する。そして、3.3 節で、都市を横断した市民意見抽出手法について説明する。

### 3.1 付与する属性の定義

本研究では、保育園と飲食店のテイクアウトサービスの2つのドメインのつぶやきを対象として市民意見抽出を行う。本節では、意見に関連する属性を複数定義するが、1つのつぶやきには複数の意見が含まれることが多い。そのため、意見に直接関連する属性 (3.1.1 節で定義する3つの属性) については、1つの意見を含む文、もしくは節を対象に属性を付与する (以降、1つの意見を含む文、もしくは節を意見ユニットと呼ぶ [11])。また、事前に意見を抽出したいと考えている話題については、あらかじめ話題との関連性を判断するモデルを訓練しておくことで、特定の話題に関連するつぶやきのみを自動抽出する。

#### 3.1.1 意見ユニットに付与する意見属性

- アプレイザル意見タイプ  
アプレイザル理論に基づき、対象に着目した意見のタイプを判断する。選択肢は「自発的感情の表明」、「人間・組織の振舞や行為を対象とした意見」、「事物・事象を対象とした意見」、「該当無し」。
- 極性  
従来の意見分析研究でも用いられている、つぶやきの極性を判断する。選択肢は「肯定」、「否定」、「中立」、「意見無し」。
- コミュニケーション意見タイプ  
モダリティおよび言語行為論を参考にして [12]、アプレイザル理論に含まれない意見タイプを判断する。これによって意見の網羅性を高める。選択肢は「推測」、「提案」、「疑問」、「要求」、「該当無し」。

#### 3.1.2 つぶやき全体に付与する属性

- 地域依存性  
意見に地域依存性があるか、すなわち、その都市で暮らす市民特有の意見か、もしくは社会一般的な意見かを判断する。地域に依存するとは、市や県等の地名や、地域の特定が可能な飲食店等の施設名を含むもの、もしくはつぶやきの内容が市民の暮らす地域に関するものである、と定義する。選択肢は「依存」、「非依存」。
- ドメインとの適合性  
この属性は、単語としては各ドメインに関連するが、

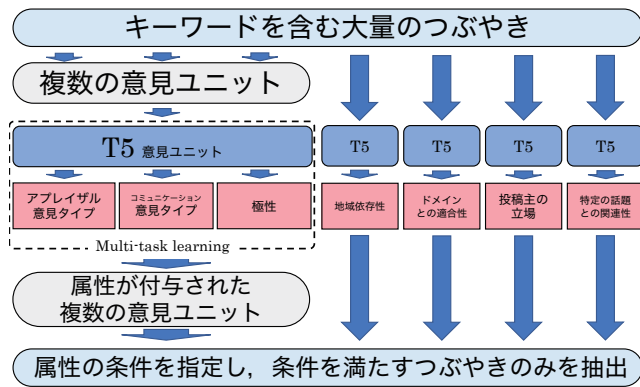


図 1 市民意見抽出のためのフレームワーク

意味的には関連のないつぶやきからの意見抽出を避けるために定義する。選択肢は「適合」、「不適合」。

● 投稿主の立場

どのようなユーザによって投稿されたつぶやきであるかを判断する。選択肢は、保育園ドメインが「小さい子を持つ親」、「保育園関係者」、「その他」、飲食店のテイクアウトサービスドメインが「店を利用した人」、「飲食店の従業員」、「その他」。

特定の話題との関連性（保育園）

● 休園・登園自粛との関連性

新型コロナウイルス感染症によって大きな問題となった、保育園の休園や登園自粛といった話題との関連性を判断する。選択肢は「関連する」、「関連しない」。

● 保育園の定員との関連性

横浜市で大きな問題となっている待機児童問題や、保育園の可否のような保育園の定員の話題との関連性を判断する。選択肢は「関連する」、「関連しない」。

特定の話題との関連性（飲食店のテイクアウトサービス）

● 商品の評価との関連性

商品の味や量、提供状態等の評価を含むかを判断する。選択肢は「関連する」、「関連しない」。

3.2 市民意見抽出のためのフレームワーク

本研究における市民意見抽出のためのフレームワークを図 1 に示す。はじめに、各ドメインに関連するキーワードを含むつぶやきを収集し、収集したつぶやきを複数の意見ユニットに分割する。意見ユニットに付与する意見属性については、T5 を用いたマルチタスク学習モデルを用いて分類を行う。意見属性は全て意見ユニット内の同じ意見についての属性であることから、これらの属性の推定タスクには関連性がある。そのため、これらの属性の分類においては、マルチタスク学習による精度向上が期待できる。つぶやき全体に付与する属性については、属性間に関連性は無いいため、独立の T5 モデルを用いて分類する。そして、これらのモデルの分類結果を属性として付与し、属性の条件

を指定することで、条件を満たすつぶやきのみを抽出する。

3.3 都市を横断した市民意見抽出

3.2 節のフレームワークを用いて、都市を横断した市民意見抽出を行う。ここでは、はじめにソース都市のつぶやきを用いて各属性の分類モデルをファインチューニングする。次に、ターゲット都市のつぶやきから、比較的少量を選択し、各属性の分類モデルを再度ファインチューニングする。この際、ターゲット都市からモデルの学習に有用な比較的少量のつぶやきを選択する必要がある。そこで、ターゲット都市のつぶやきをランダムに選択する場合と、ソース都市でファインチューニングしたモデルによる予測の確信度が高いつぶやきを選択する場合との効果を比較し、ターゲット都市のつぶやきに対するアノテーションのコストを効果的に削減できるか検証する。

4. 実験

本節では、市民意見抽出のフレームワークを用いた、都市を横断した各属性の分類実験について述べる。4.1 節では、本研究で使用するつぶやきの収集方法と、市民意見分析コーパスの人手によるアノテーションについて述べ、4.2 節で実験の方法、目的、結果について述べる。

4.1 市民意見分析コーパス

はじめに、プロフィール情報に基づいて収集した横浜市民の Twitter アカウント 82,583 件と札幌市民 64,790 件のつぶやきを、Twitter の Streaming API を用いて収集した。これらのつぶやきのうち、2020 年 1 月 1 日から 2020 年 7 月 11 日までの横浜市民の計 28,971,414 件のつぶやきと、札幌市民の計 19,286,694 件のつぶやきから、「保育園」、「保育士」、「保活」、「待機児童」の単語を含むつぶやきを保育園ドメインのつぶやきとして収集し、「持ち帰り」、「テイクアウト」の単語を含むつぶやきを飲食店のテイクアウトサービスドメインのつぶやきとして収集した。リツイートや重複するつぶやきは取り除いた。

本研究で用いるデータ数は両都市の両ドメインに共通で 2,622 件のつぶやきと、それらを文単位に区切り直したものとす。つぶやきの文単位への区切り直しには、Python のライブラリ、spaCy\*2を用いた。この際、名詞のみで構成される文は意見性を含まないものが多いため、spaCy による区切り直し後に名詞のみで構成された文が抽出された場合、つぶやきの先頭以外に現れる文は 1 つ前の文に結合し、先頭に現れる文は 1 つ後ろの文に結合する処理を行った。また、改行は文の区切りとした。さらに、ハッシュタグのみの文も意見性を含まないものが多いため、前の文と結合する処理を行った。最後に、閉じ括弧から始まる文につ

\*2 <https://spacy.io/>

いては、前の文と結合する処理を行った。

収集した横浜市民と札幌市民のつぶやきに 3.1 節で定義した各属性を付与するアノテーション作業を人手で行い、意見分析コーパスを作成した。アノテーション作業は、各都市ごとに第一著者を含む合計 5 名の判定者によって行い、全てのアノテーション結果は多数決によって決定した。

アノテーション作業では、はじめに各判定者のアノテーション方針を一致させるための訓練を行った。訓練では、保育園ドメイン 250 文と飲食店のテイクアウトサービスドメイン 250 文の計 500 文と、各文を含むつぶやきを対象に、5 名全員で各属性を判定し、判定者のアノテーション方針が一致した時点で訓練を終了した。この際、1 文に複数の意見が含まれると判断した場合は、複数の意見が現れない意見ユニットとなるまで分割した。分割作業は全員で意見を交換し、過半数の判定者間で一致した文でのみ行った。

訓練終了後に、残りの全てのアノテーション作業を行った。第一著者は全てのつぶやき、文のアノテーションを行い、両都市において、残りの 4 名を 2 名ずつに分けることで、各 3 名ずつの 2 チームを作り、各チームが半数ずつを担当した。この際も、1 文に複数の意見が含まれると過半数の判定者間で判断が一致した場合は、複数の意見が現れない意見ユニットとなるまで分割した。Fleiss の  $\kappa$  係数 [13] を用いた各チームのアノテーションの一致度は、両都市の全ての属性において 0.6 (Substantial Agreement[14]) 以上となり、判定者によって属性の判定に大きな差異が生まれないことが示された。

アノテーション結果が 1-1-1 や 2-2-1 となり、多数決によって結果を決められない場合については、判定者間で意見を交換することで、全ての結果を多数決で決定した。

最終的に得られた市民意見分析コーパスの各都市におけるつぶやきと意見ユニットのデータ数を表 1 に示す。

| 都市  | ドメイン   | つぶやき  | 意見ユニット |
|-----|--------|-------|--------|
| 横浜市 | 保育園    | 2,622 | 7,916  |
|     | テイクアウト | 2,622 | 6,671  |
| 札幌市 | 保育園    | 2,622 | 8,113  |
|     | テイクアウト | 2,622 | 7,620  |

## 4.2 都市を横断した各属性の分類実験

### 4.2.1 実験の目的

本実験では、横浜市と札幌市の市民のつぶやきからなる市民意見分析コーパスを用いて、3.3 節の手法による都市を横断した各属性の分類を行い、手法の有効性を検証する。

### 4.2.2 実験の方法

本実験では、はじめにソース都市の市民意見分析コーパスの全データを用いて、各属性の分類モデルをファインチューニングする。次に、ターゲット都市のデータをラン

ダムに 0%, 25%, 50% と一部用いることで、ソース都市で訓練済みのモデルを再度ファインチューニングする。なお、ターゲット都市のデータが 0% というのは、ソース都市のデータのみでファインチューニングしたモデルを用いてターゲット都市における各属性の分類を行うことを表しているため、2 段階のファインチューニングは行わない。

意見ユニットに付与する属性の分類においては、各属性を独立の T5 モデルで分類する独立学習の手法を比較手法として、各属性を 1 つの T5 モデルで分類するマルチタスク学習手法の有効性を検証する。一方で、つぶやき全体に付与する属性では、各属性間に関連性は無いいため、全ての分類モデルは独立の T5 モデルを用いており、比較手法として、ターゲット都市のデータのみを用いてモデルをファインチューニングした際の精度を算出する。

また、ターゲット都市から一部使用するデータについて、ランダムに選択する手法と、モデルの予測の確信度が高いデータを選択する手法の比較を行う。モデルの予測の確信度が高いデータの選択では、はじめにソース都市のデータを使って T5 モデルをファインチューニングし、ターゲット都市のデータの各属性を推定する。そして、ターゲット都市のデータのうち、モデルの予測の確信度\*3が高いデータから順に一部用いることで、2 段階目のファインチューニングを行い、ターゲット都市の各属性の予測精度を算出する。この比較実験では、マルチタスク学習を用いた意見属性の推定において、50% のターゲット都市のデータを用いる場合について検証する。

本実験の評価指標は F 値とし、5 分割交差検証を用いてコーパス内の全てのデータの各属性の分類精度を算出する。5 分割交差検証の各検証では、全体の 20% のデータをテストデータ、16% のデータを検証データ、64% のデータを訓練データとして用いる。ターゲット都市のデータが 0% の場合は、訓練データとテストデータが完全に別のデータとなるため、5 分割交差検証は行わない。5 分割交差検証では、ターゲット都市のデータのみでモデルをファインチューニングした際のテストデータと、ターゲット都市のデータを 25%, 50% 用いる場合のテストデータは全て同じものであり、ターゲット都市のデータを 25%, 50% 用いる場合は、モデルのファインチューニングに使用する訓練データ、検証データの割合のみを変化させる。

本実験では、Python の Hugging Face transformers ライブラリの事前学習済み T5 モデル\*4を用いてファインチューニングを行う。ファインチューニングのエポック数は、5 分割交差検証の各検証において、2,3,4,5 エポックのうち、検証データにおける F 値が最も高いものを用いる。

\*3 T5 モデルの予測の際に使用しているビームサーチのスコア (Hugging Face 中ではパラメタ `sequences_scores` と定義) のうち、予測ラベルに対応するスコアを確信度とする

\*4 <https://huggingface.co/sonoisa/t5-base-japanese>

表 2 都市を横断した意見ユニットの属性の分類精度 (F 値)

| ドメイン   | 属性                 | モデル          | 横浜市を評価                          |              |              |                   | 札幌市を評価                          |              |              |                   |
|--------|--------------------|--------------|---------------------------------|--------------|--------------|-------------------|---------------------------------|--------------|--------------|-------------------|
|        |                    |              | 札幌市で訓練し<br>横浜市の訓練データを<br>追加した割合 |              |              | 横浜市のみで訓練<br>(上限値) | 横浜市で訓練し<br>札幌市の訓練データを<br>追加した割合 |              |              | 札幌市のみで訓練<br>(上限値) |
|        |                    |              | 0%                              | 25%          | 50%          |                   | 0%                              | 25%          | 50%          |                   |
| 保育園    | アプレイザル<br>意見タイプ    | マルチタスク       | 0.560                           | <b>0.635</b> | <b>0.665</b> | <b>0.701</b>      | <b>0.568</b>                    | <b>0.661</b> | <b>0.687</b> | <b>0.705</b>      |
|        |                    | 独立学習         | <b>0.580</b>                    | 0.630        | 0.642        | 0.636             | 0.567                           | 0.648        | 0.652        | 0.651             |
|        | 極性                 | マルチタスク       | 0.730                           | <b>0.793</b> | <b>0.819</b> | <b>0.849</b>      | <b>0.752</b>                    | <b>0.795</b> | <b>0.826</b> | <b>0.864</b>      |
|        |                    | 独立学習         | <b>0.746</b>                    | 0.780        | 0.790        | 0.792             | 0.750                           | 0.775        | 0.788        | 0.774             |
|        | コミュニケーション<br>意見タイプ | マルチタスク       | 0.719                           | <b>0.734</b> | <b>0.759</b> | <b>0.764</b>      | 0.729                           | 0.736        | 0.755        | <b>0.764</b>      |
|        |                    | 独立学習         | <b>0.737</b>                    | 0.724        | 0.743        | 0.739             | <b>0.734</b>                    | <b>0.754</b> | <b>0.758</b> | 0.754             |
| テイクアウト | アプレイザル<br>意見タイプ    | マルチタスク       | 0.628                           | <b>0.722</b> | <b>0.731</b> | <b>0.771</b>      | 0.616                           | <b>0.706</b> | <b>0.736</b> | <b>0.770</b>      |
|        |                    | 独立学習         | <b>0.634</b>                    | 0.700        | 0.704        | 0.698             | <b>0.618</b>                    | 0.690        | 0.711        | 0.709             |
|        | 極性                 | マルチタスク       | 0.793                           | <b>0.816</b> | <b>0.841</b> | <b>0.869</b>      | 0.774                           | <b>0.810</b> | <b>0.833</b> | <b>0.877</b>      |
|        |                    | 独立学習         | <b>0.794</b>                    | 0.803        | 0.809        | 0.807             | <b>0.775</b>                    | 0.790        | 0.803        | 0.805             |
|        | コミュニケーション<br>意見タイプ | マルチタスク       | 0.736                           | 0.757        | <b>0.783</b> | <b>0.764</b>      | 0.680                           | 0.748        | 0.769        | <b>0.793</b>      |
|        |                    | 独立学習         | <b>0.737</b>                    | <b>0.769</b> | 0.780        | 0.760             | <b>0.695</b>                    | <b>0.777</b> | <b>0.789</b> | 0.786             |
| 平均     | マルチタスク             | 0.685        | <b>0.743</b>                    | <b>0.766</b> | <b>0.786</b> | 0.687             | <b>0.743</b>                    | <b>0.768</b> | 0.796        |                   |
|        | 独立学習               | <b>0.745</b> | 0.734                           | 0.745        | 0.739        | <b>0.690</b>      | 0.739                           | 0.750        | 0.747        |                   |

注) 2行ごとに各属性の分類におけるマルチタスク学習手法と独立学習手法の精度を比較し、高精度な手法を太字で強調

表 3 都市を横断したつぶやきの属性の分類精度 (F 値)

| ドメイン   | 属性           | 横浜市を評価                          |              |              |          | 札幌市を評価                          |              |              |          |
|--------|--------------|---------------------------------|--------------|--------------|----------|---------------------------------|--------------|--------------|----------|
|        |              | 札幌市で訓練し<br>横浜市の訓練データを<br>追加した割合 |              |              | 横浜市のみで訓練 | 横浜市で訓練し<br>札幌市の訓練データを<br>追加した割合 |              |              | 札幌市のみで訓練 |
|        |              | 0%                              | 25%          | 50%          |          | 0%                              | 25%          | 50%          |          |
| 保育園    | 地域依存性        | 0.892                           | 0.924        | <b>0.927</b> | 0.925    | <b>0.885</b>                    | <b>0.891</b> | <b>0.888</b> | 0.884    |
|        | ドメインとの適合性    | 0.897                           | <b>0.913</b> | <b>0.915</b> | 0.909    | 0.839                           | 0.850        | 0.861        | 0.862    |
|        | 休園・登園自粛との関連性 | 0.785                           | 0.837        | <b>0.855</b> | 0.854    | 0.785                           | 0.790        | <b>0.811</b> | 0.795    |
|        | 定員との関連性      | <b>0.859</b>                    | <b>0.862</b> | 0.845        | 0.850    | <b>0.774</b>                    | <b>0.750</b> | <b>0.777</b> | 0.728    |
|        | 投稿主の立場       | 0.600                           | 0.658        | 0.702        | 0.709    | <b>0.667</b>                    | 0.623        | <b>0.660</b> | 0.648    |
| テイクアウト | 地域依存性        | 0.889                           | 0.904        | 0.917        | 0.926    | 0.871                           | 0.893        | 0.901        | 0.906    |
|        | ドメインとの適合性    | 0.984                           | 0.983        | 0.984        | 0.987    | 0.982                           | <b>0.985</b> | <b>0.985</b> | 0.985    |
|        | 商品の評価との関連性   | 0.832                           | 0.850        | 0.850        | 0.855    | 0.822                           | 0.836        | 0.847        | 0.853    |
|        | 投稿主の立場       | 0.851                           | 0.874        | 0.877        | 0.885    | 0.882                           | <b>0.892</b> | <b>0.898</b> | 0.888    |

注) ターゲット都市のみでファインチューニングした際の精度以上の値を太字で強調

#### 4.2.3 実験結果

##### 意見ユニットに付与する各属性の分類

都市を横断した、意見ユニットに付与する各属性の分類精度を表 2 に示す。実験結果から、ターゲット都市のデータのみを用いてモデルを訓練した場合<sup>\*5</sup>、横浜市、札幌市の全ての属性において、マルチタスク学習モデル (平均 = 0.791, 標準偏差 = 0.061) が独立学習モデル (平均 = 0.743, 標準偏差 = 0.058) と比較して高い F 値となり、Pearson's  $r$  を効果量とした 1 対の対応のある両側  $t$  検定で有意差が確認できた ( $t(11) = 5.750, p = 0.0001, \text{Pearson's } r = 0.866$ )。また、ターゲット都市のデータを 25% 用いた場合<sup>\*6</sup>においては、横浜市で評価した際のテイクアウトドメインのコミュニケーション意見タイプと札幌市で評価した際のコミュニケーション意見タイプを除いた全ての属性においてマルチタスク学習モデルが独立学習モデルと比較して高い F 値となっているが、2つの手法に有意差は確認で

きなかった。さらに、ターゲット都市のデータを 50% 用いた場合<sup>\*7</sup>においても、札幌市で評価した際のコミュニケーション意見タイプを除いた全ての属性においてマルチタスク学習モデル (平均 = 0.767, 標準偏差 = 0.057) が独立学習モデル (平均 = 0.747, 標準偏差 = 0.058) と比較して高い F 値となっており、Pearson's  $r$  を効果量とした 1 対の対応のある両側  $t$  検定で有意差が確認できた ( $t(11) = 3.865, p = 0.002, \text{Pearson's } r = 0.759$ )。以上の結果より、本研究における意見ユニットに付与する属性の分類には、各属性間の依存関係を学習するマルチタスク学習モデルの適用が有効であることが分かる。

ターゲット都市のデータの選択については、モデルの予測の確信度上位のデータを選択する手法 (平均 = 0.793, 標準偏差 = 0.063) が、ランダムにデータを選択する手法 (平均 = 0.767, 標準偏差 = 0.057) と比較して高精度な結果となり<sup>\*8</sup>, Pearson's  $r$  を効果量とした 1 対の対応のある両側  $t$

\*5 表 2 の「横浜市のみで訓練」、「札幌市のみで訓練」の列に該当

\*6 表 2 の「25%」の列に該当

\*7 表 2 の「50%」の列に該当

\*8 全ての実験結果の値は、紙面のスペースの都合により省略する

検定で有意差が確認できた ( $t(11) = -0.674, p = 0.00003$ , Pearson's  $r = 0.896$ ). そのため, 新たにターゲット都市のデータをアノテーションする場合は, ソース都市のデータで訓練したモデルによる予測の確信度が高いデータから順に選択することで, アノテーションのコストを効果的に削減できると分かる. さらに, これらの結果は, 全ての属性で, ターゲット都市のデータのみを用いた独立学習モデル以上の精度となっていることから, 新たなターゲット都市での市民意見分析を行う場合, ソース都市の 50% (約 1,300 件分のつぶやきに該当) のデータをアノテーションすれば, 提案手法により一定の精度が実現できるといえる.

### つぶやき全体に付与する各属性の分類

都市を横断した, つぶやき全体に付与する各属性の分類精度を表 3 に示す. 表より, 多くの属性において, ターゲット都市のデータを一部用いることで, ターゲット都市のみでファインチューニングした際の精度以上の精度となっていることが分かる. また, 横浜市を評価する際のテイクアウトドメインの各属性など, ターゲット都市のみでファインチューニングした際の精度以上とはならなかった属性においても, 全ての属性で, ターゲット都市のデータを一部用いることで, ターゲット都市のみでファインチューニングした際の精度と 0.01 以内の F 値の差の精度で分類が可能となっている. これらの結果から, つぶやき全体に付与する属性は, ターゲット都市のデータを一部用いることで, ターゲット都市のみでファインチューニングした際の精度と同程度かそれ以上の精度で分類できることが分かった.

## 5. おわりに

本研究では, 著者らの先行研究で提案した特定の都市における市民意見抽出のためのフレームワークを拡張し, 都市を横断した市民意見抽出のための手法の検証を行った. 実験では, 横浜市民と札幌市民のつぶやきにアノテーションをすることで, 市民意見分析コーパスを作成し, ソース都市と一部のターゲット都市のデータを用いて, 各属性の分類モデルを 2 回ファインチューニングする手法の有効性を検証した. つぶやき全体に付与する属性については, ターゲット都市のデータを一部用いることで, ターゲット都市のみでファインチューニングした際の精度と近い精度か, それ以上の精度で各属性が分類可能であることを示した. 意見ユニットに付与する意見属性については, ターゲット都市のデータを一部用いてマルチタスク学習を適用することで, 多くの属性においてターゲット都市のデータのみで各属性を独立にファインチューニングした際の精度を上回ることが分かり, 提案手法の有効性を確認した.

また, 表 2 より, ターゲット都市のデータを使わない場合<sup>\*9</sup>は, 多くの属性で独立学習手法がマルチタスク学習手

法を上回る結果となった. そのため, 今後はソース都市において各属性を独立に学習した情報をもとに, 新たにターゲット都市のデータを用いてマルチタスク学習モデルを構築するアルゴリズムの検討を行いたい.

**謝辞** 本研究の一部は, 科学研究費補助金基盤研究 B (課題番号 19H04420), 挑戦的研究 (萌芽) (課題番号 22K19822), 2022 年度国立情報学研究所公募型共同研究 (採択番号 22S0103) の助成を受けて遂行された.

### 参考文献

- [1] Ishida, T., Seki, Y., Kashino, W. and Kando, N.: Extracting Citizen Feedback from Social Media by Appraisal Opinion Type Viewpoint, *Journal of Natural Language Processing*, Vol. 29, No. 2, pp. 416–442 (2022).
- [2] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*.
- [3] Martin, J. R. and White, P. R. R.: *The Language of Evaluation: Appraisal in English*, Palgrave Macmillan (2005).
- [4] Halliday, M. A. K.: *An Introduction to Functional Grammar*, Edward Arnold (1985).
- [5] 関 洋平: コミュニティ QA における意見分析のためのアノテーションに関する一検討, *自然言語処理*, Vol. 21, No. 2, pp. 271–299 (2014).
- [6] Mora, N. and Lavid-Lopez, J.: Building an Annotated Aataset of App Store Reviews with Appraisal Features in English and Spanish, *Proc. of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pp. 16–24 (2018).
- [7] Caruana, R.: Multitask Learning, *Machine Learning*, Vol. 28, pp. 41–75 (1997).
- [8] Mitchell, T. M.: The Need for Biases in Learning Generalizations, Technical Report CBM-TR-117, Rutgers University, New Brunswick, NJ (1980).
- [9] Liu, X., He, P., Chen, W. and Gao, J.: Multi-Task Deep Neural Networks for Natural Language Understanding, *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4487–4496 (2019).
- [10] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P. J.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, *Journal of Machine Learning Research*, Vol. 21, No. 140, pp. 1–67 (2020).
- [11] Seki, Y., Ku, L. W., Sun, L., Chen, H. H. and Kando, N.: Overview of Multilingual Opinion Analysis Task at NTCIR-8 - A Step Toward Cross Lingual Opinion Analysis, *Proc. of the Eighth NTCIR Workshop Meeting*, pp. 209–220 (2010).
- [12] 大塚裕子, 乾 孝司, 奥村 学: 意見分析エンジン - 計算言語学と社会学の接点-, コロナ社 (2007).
- [13] Fleiss, J. L.: Measuring Nominal Scale Agreement Among Many Raters, *Psychological Bulletin*, Vol. 76, No. 5, pp. 378–382 (1977).
- [14] Landis, J. R. and Koch, G. G.: The Measurement of Observer Agreement for Categorical Data, *Biometrics*, Vol. 33, No. 1, pp. 159–174 (1977).

\*9 表 2 の「0%」の列に該当