## 概要

ソーシャルメディアの発展により，多くの新しい情報がユーザによって発信されている．これに伴い，トレンドワード検出やオピニオンマイニングなど，ソーシャルメディア上の情報の分析の重要性が増している．この際，性別，年齢，出身地など，ユーザのプロフィールを考慮することで，より詳細な分析が可能となる．ただし，代表的なソーシャルメディアのひとつである Twitter では，一般にユーザのプロフィールは公開されていない．そのため，Twitter ユーザのプロフィールを高い精度で推測する技術が求められる．先行研究では，ユーザのツイートからプロフィールを予測する手法が提案されているが，ツイートのテキストのみを素性とする手法が主流であり，テキストおよび投稿画像の両方を同時に考慮する手法は十分に研究されていない．また，自己紹介文やフォロー数など，ツイート以外の情報を素性として用いる研究もあるが，使用される情報は限られている．

本研究では，プロフィールの中でも性別に着目し，教師あり学習によって Twitter ユーザの性別を自動推定する手法を提案する．具体的には，ツイートのテキストと画像を同時に素性として用いるモデル (分類器) を提案する．さらに，ツイートを含め，Twitter から得られる多様な情報を考慮した性別推定モデルを提案する．また，訓練データを自動的に拡張することで性別推定の正解率を向上させることを試みる．

まず，Twitter ユーザに対して性別のラベルを自動的に付与し，性別推定モデルを学習するための訓練データを構築する．著名人は性別を推測しやすいと考え，フォロー数の多いユーザを著名人とみなして収集し，Wikipedia などの外部情報や顔画像認識ツールを用いて，その性別を推定する．結果として，およそ 5,000 名のユーザに対して性別ラベルを付与したデータセットを構築した．以下，これを「初期データ」と呼ぶ．

次に，性別推定の分類器を学習する手法について述べる．最初に，ユーザが投稿した複数のツイートから性別を予測する分類器を学習する．まず，それぞれのツイートに対して，その性別スコアを算出する．性別スコアとは，0〜1 までの数値で，0 は女性，1 は男性であることを表す．そして，複数のツイートの性別スコアの平均により，ユーザの性別スコアを決定する．ツイートの性別を予測するモデルとして，投稿テキストと投稿画像を同時に考慮した複数の分類器を含め，以下の 7 つの分類器を提案する．Text only model は，ツイートのテキストのみを利用し，Bidirectional Encoder Representations from Transformers (BERT) を用いたモデルである．Image only model は，ツイートの画像のみを利用し，Vision Transformer (VT) をベースとしたモデルである．Early fusion model は，テキストを BERT，画像を VT で埋め込み表現に変換し，これを結合して全結合層に渡して，性別スコアを予測するモデルである．Late fusion model では，同じくテキストと画像の埋め込み表現を得た後，それぞれを全結合層で 4 次元のベクトルに縮退する．これらを結合して別の全結合層に渡して，性別スコアを予測する．Dense fusion model は，Early fusion model と Late fusion model を組み合わせたモデル

である．Caption model は，既存のツールで画像のキャプションを生成し，これをツイートのテキストと連結した文を入力とし，BERT を用いて性別を予測するモデルである．Ensemble model では，画像付きのツイートは Early fusion model，画像のないツイートは Text only model を用いて性別を推定する．

次に，Twitter から得られる様々な情報を用いて 6 種類の分類器を作成する．1 つ目は前述のツイートを用いた分類器である．2 つ目は，フォロワー数や投稿数など，Twitter 固有の統計量を素性とし，Light Gradient Boosting Machine (Light GBM) を用いて学習した分類器である．3 つ目は，ユーザ名の文字 uni-gram などを素性とした Light GBM である．4 つ目は，プロフィール画像，ヘッダー画像 (Twitter のユーザページの上部に表示される画像)，もしくは両方を素性とし，VT で学習された分類器である．5 つ目は，ユーザの自己紹介文を素性とし，BERT で学習された分類器である．6 つ目は，ユーザのフォロイーの自己紹介文から，bag-of-words による素性ベクトルを作成し，これを入力とする Light GBM である．さらに，これらの分類器が出力する性別スコアを素性とし，Light GBM によって性別を推測する「統合モデル」を提案する．

初期データは著名人のユーザから構成されているが，これから学習した性別推定モデルは，一般ユーザに対する性別推定の性能が低下すると考えられる．この問題を解決するため，一般ユーザに対して自動ラベル付けを行い，訓練データを拡張する．これを「データ拡張」と呼ぶ．まず，一般ユーザのアカウントを収集し，ラベルなしデータを作成する．初期データで学習した性別推定の分類器を用いて，ラベルなしデータのユーザの性別を推測し，判定の信頼度の高い上位 3,000 件のユーザを選択し，訓練データに追加する．

提案手法の評価実験を行った．人手評価データとして，フォロワー数が 1,000 人以下の 459 名の一般ユーザに対し，3 名の被験者によって性別ラベルを付与した．ツイートを用いた分類器の評価では，初期データを訓練データとテストデータに分割する実験と，初期データを訓練データ，人手評価データをテストデータとする実験を行った．

ツイートを用いた分類器の性別推定の正解率を比較したところ，初期データを用いた実験では Early fusion model の正解率が一番高く，0.893 であった．人手評価データを用いた実験では，Caption model の正解率が一番高く，0.789 であった．次に，人手評価データをテストデータとし，ツイート以外の情報を用いた分類器ならびに統合モデルを評価した．その結果，プロフィール画像，ユーザ名，ツイートを用いた分類器の正解率が比較的高かった．最高の正解率が得られたのは統合モデルで，その正解率は 0.851 であった．

データ拡張の評価では，初期データのみで訓練した分類器と，データ拡張によって得られたデータを訓練データに加えて学習した分類器の正解率を比較した．データ拡張の際，評価対象の分類器を用いて自動ラベル付けを行う場合と，統合モデルを用いて自動ラベル付けを行う場合を比較した．後者では，質の高い訓練データ量の増加によって正解率がどれだけ向上するかを検証した．実験の結果，5 つの

分類器について，データ拡張によって正解率が向上した．ただし，統合モデルでは正解率が改善しなかった．また，単独の分類器よりも統合モデルを用いてデータ拡張をした方が，データ拡張後の正解率が高くなる傾向が確認された．ただし，データ拡張後の分類器はいずれも，データ拡張を行わない統合モデルの正解率を越えることはなく，データ拡張のはっきりとした有効性は確認できなかった．