修士論文

多様な特徴量を同時に考慮したソーシャルメディア上のユーザプロフィール推定

廣田 遼

主指導教員 白井 清昭

北陸先端科学技術大学院大学
先端科学技術研究科
(情報学科)

令和 5 年 3 月

Abstract

With the development of social media, the Internet is flooded with new information released by users. Therefore, analysis of information on social media, such as finding trending words and opinion mining, becomes more important. More detailed analysis is possible by considering users' profiles such as a gender, age, hometown, etc. However, in one of the most poplar social media, Twitter, the profiles of the users are not made in public. Therefore, a technique to precisely predict the profiles of Twitter users is required. Although previous work on the prediction of users' profiles mainly focuses on texts of tweets posted by the users, methods considering both texts and images in tweets are not paid much attention. Several previous studies also use information other than tweets, such as a text of self-introduction by the user and the number of followers, but only the limited information has been utilized.

This thesis focuses on gender among various information of user profiles and proposes novel methods based on supervised learning for prediction of gender of Twitter users. Specifically, models or classifiers that consider texts and images of tweets as features simultaneously are proposed. Models of gender prediction using a wide variety of information obtained from Twitter, including tweets, are also proposed. Furthermore, the size of training data is increased automatically to improve the performance of gender prediction.

First, training data for supervised learning of a gender prediction model is constructed by automatically determining gender of Twitter users. Supposing that celebrities are relatively easy to guess their gender, users who have many followers are retrieved as celebrities, then their gender are predicted using publicly available information such as Wikipedia and a face image recognition tool. As a result, a dataset consisting of approximately 5,000 users with their gender labels is constructed. This dataset is called "initial labeled data".

Several kinds of classifiers for gender prediction are developed. The first type is a tweet-based classifier that predicts gender from multiple tweets posted by the user. First, a gender score for each tweet is calculated. Here, the gender score is defined as a value between 0 and 1, where 0 and 1 indicate female and man respectively. Then, the average of the gender scores of multiple tweets is used to calculate the gender score of the user. To predict the gender of each tweet, the following seven classifiers are proposed, including classifiers that simultaneously consider both the text and image in the tweet. "Text only model" uses only the text of the tweet and is based on Bidirectional Encoder Representations from Transformers (BERT). "Image only model" uses only the image of the tweet and is based on the Vision Transformer (VT). "Early fusion model" uses BERT and VT to obtain embedding of the text and image respectively, then those embeddings are

concatenated and passed to the Fully Connected Layer (FCL) to obtain the gender score. "Late fusion model" obtains the text and image embeddings in the same way, and converts each to a 4-dimensional vector by FCL. Then, those vectors are concatenated and passed to another FCL to obtain the gender score. "Dense fusion model" is a classifier that combines the Early and Late fusion models. "Caption model" first uses an existing tool to generate a caption of the image in the tweet. Then, BERT is applied where the text of the tweet and the generated image caption are given as an input. "Ensemble model" uses the Early fusion model when the target tweet includes an image, otherwise uses the Text only model.

In addition, six classifiers using various information obtained from Twitter are proposed. The first classifier is the tweet-based classifier described earlier. The second classifier is one that trained by Light Gradient Boosting Machine (Light GBM) using Twitter-specific statistics such as the number of followers and posts as features. The third classifier considers a name of the Twitter user. A Light GBM is trained using character uni-grams of the user name. The fourth classifier is one using a profile image of the user, a header image shown in the user page on Twitter, or both, where VT is used as the base model. The fifth classifier is a BERT model using the text of self-introduction written by the user as a feature. The sixth classifier considers the texts of self-introduction of followees. Bag-of-words features are extracted from those texts, then a Light GBM is training with them. Furthermore, "Integrated model" is proposed. It uses the gender scores obtained by the above classifiers as features and predicts gender of the user by a Light GBM.

Since the initial labeled data consists of users of celebrities, the gender prediction model trained from it is expected to perform poorly for general users. To solve this problem, the training data is expanded by automatically labeling general users. This method is called "data augmentation". First, unlabeled data is made by retrieving Twitter accounts of general users. Then, the gender of the users in this data is predicted by the classifier trained from the initial labeled data. The top 3,000 users whose gender is the most reliably predicted by the model are chosen, and added to the training data.

Several experiments were conducted to evaluate the proposed methods. First, "manually labeled data" was constructed, which was used as the test data. Three human subjects determined the gender of 459 general users who had less than 1,000 followers. Then, the proposed methods were evaluated in two experimental settings. One is that the initial labeled data was divided into the training, development and test data, the other is that the initial labeled data was used as the training data and the manually labeled data as the test data.

The tweet-based classifiers were firstly evaluated. Early fusion model achieved

the highest accuracy of 0.893 in the experiment using the initial labeled data. On the other hand, the accuracy of Caption model was the highest, 0.789, in the experiment using the manually labeled data. Next, the performance of the individual classifiers and the Integrated model using the information other than tweets were evaluated using the manually labeled data as the test data. The results showed that the classifiers using the profile image, user name, and tweets achieved relatively high accuracy. The highest accuracy, 0.851, was obtained by the Integrated model.

In the evaluation of data augmentation, the accuracy of the classifiers trained with only the initial labeled data was compared with the classifiers trained with both the augmented and initial training data. Two methods of the data augmentation were compared: one was that the same classifier was used to determine the gender of unlabeled data to expand the data to train itself, the other was that Integrated model was used for determination of gender. In the latter case, it was mainly assessed how much the increase of high-quality training data could improve the accuracy of gender prediction. Experimental results showed that data augmentation could improve the accuracy for five classifiers. However, the performance of Integrated model was not improved by data augmentation. It was found that the performance of the classifiers tended to be improved by the data augmentation using the Integrated model than that using the same classifier. However, none of the classifiers with data augmentation did not outperform Integrated model without data augmentation. Thus the effectiveness of data augmentation was not clearly confirmed.

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．

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　背景

近年ではインターネットの普及によって様々なサービスが発展している．これに伴い，メディアのあり方にも変化が訪れている．以前の代表的なメディアはテレビ，新聞，ラジオのようなマスメディアであった．しかしソーシャルメディアの発展は多くの人々の生活に影響を与えた．例えば，テレビを持たない若者がいたり，新聞の発行部数が減少したりするなど，ソーシャルメディアや電子サービスがマスメディアの代わりを担っていると感じることも多くなった．ソーシャルメディアの特長として，双方向のやり取りができること，個人が少ないコストで世界中に情報発信することができること，これにより誰もが注目を集めることが容易になったこと，などが挙げられる．近年では，なりたい職業ランキングにインフルエンサーがランクインすることもあるように，ソーシャルメディアは価値観の変化を生んでいる．また，ソーシャルメディアがワールドカップでの盛り上がりの一端を担うことや，政治活動の情報発信の場になることもある．

このように多くのユーザが注目するソーシャルメディアは個人だけでなく企業にとっても利用価値の高いメディアの１つである．自社の商品やサービスに関する情報発信はもちろん，ユーザの意見を把握したり，ユーザとの双方向のやり取りを行ったり，ユーザの流行を把握したりするなど，活用の幅は広い．本論文は，代表的なソーシャルメディアのひとつである Twitter に注目する．Twitter の特徴として，発信する情報は主にテキストと画像であること，情報のリアルタイム性が強いこと，拡散力が高いことが挙げられる．Twitter では毎日多くのユーザが情報を発信し，注目されるトレンドの移り変わりが早く，全ての流行を把握することは難しい．このような日々更新されていく大量の情報を整理する方法として，トレンドワード検出やオピニオンマイニングなど，Twitter に投稿されたツイートを対象としたデータ分析の重要性が増している．特に，性別や年齢といったユーザのプロフィール情報を考慮した分析の需要は高い．例えば，男性と女性とで人気にある商品に違いはあるか，若年層と老年層で流行っている言葉に違いはあるか，といった分析結果は，企業やサービスプロバイダにとって，製品やサービス提供のヒントに直接つながる情報であり，価値が高い．また，Twitter は一般に公開されているので競合他社の製品やサービスに対する評判を分析することも可能であり，自社との違いを理解することで今後の経営の方針を決定するといった活用も考えられる．

しかし，Twitter では，性別，年齢，出身地などといったユーザのプロフィールの情報は本人が開示しない限り公開されていない．そのため，現時点では，プロフィールを開示してユーザのみを対象に調査を実施したり，公開されている情報からある程度ユーザのプロファイルを推測することで，前述のようなプロフィール情報を考慮した情報分析を行う必要がある．

1.2 目的

上記の研究背景を踏まえ，本研究では，プロフィールのうち性別に着目し，Twitter ユーザの性別を自動推定する新しい手法を提案する．特に，プロフィール情報が開示されていることの多い限られた著名人ではなく，プロフィール情報を開示していない一般ユーザのプロフィールを自動推定できる手法を提案する．本研究におけるプロフィール推定は教師あり学習に基づく．学習に用いる素性は，Twitter から一般に取得できる多様な情報である．具体的には，ツイート，Twitter 固有の統計量，ユーザ名，プロフィール画像，ヘッダー画像，自己紹介文，ユーザがフォローしている人たちの自己紹介文である．これらのそれぞれを素性とする複数の性別推定モデルを学習する．さらに，これらのモデルの結果を素性とし, 多様な情報を総合的に評価して性別を予測するモデルも学習する．しかし，教師あり学習のためには性別ラベルが付与されたデータセットが必要だが，Twitter による個人情報を保護する規定からユーザプロフィールは一般に公開されていないため，このようなデータセットを自動構築することは難しい．一方，人手によるデータセットの作成には多大なコストがかかる．そこで，フォロー数の多い主に著名人のユーザを対象に，Wikipedia などの外部情報を用いて性別の自動ラベル付けを行う．ただし，著名人のデータから学習された性別判定モデルは，ドメインシフトにより，一般のユーザに対する性別判定の性能が下がることが予想される．そのため，著名人のデータセットで学習したモデルを用いて，ラベルが付与されていない一般ユーザに対して性別判定を行い，自動ラベル付けによるデータ拡張を行うことで，ドメインシフトの問題を軽減することも狙う．

1.3 本論文の構成

本論文は 5 つの章から構成される．第 2 章では Twitter 情報を用いたプロフィール予測の先行研究を紹介する．また，先行研究に対する本研究の特徴も論じる．第 3 章では，本研究で提案する Twitter 情報からユーザの性別を予測するモデルを詳述する．第 4 章では，提案手法の評価実験について報告する．最後に第 5 章では，本論文のまとめと今後の課題について述べる．

第2章　関連研究

本章では，本論文の関連研究について述べる．2.1 節では，Twitter ユーザのプロフィールを予測する過去の研究を紹介する．2.2 節では，提案手法の中で利用する言語モデルや機械学習モデルを紹介する．2.3 節では，先行研究と本研究の違いについて論じる．

2.1　Twitter ユーザのプロフィール予測に関する研究

2.1.1　性別予測に関する研究

Sakaki らは，Twitter ユーザの性別を推定するために，ユーザが投稿したツイートのテキストならびに画像を主な素性として，性別判定の分類器を学習する手法を提案した [11]．従来研究の多くはツイートのテキストのみを用いて性別を推定する分類器を学習する研究が多かったのに対し，ツイートの投稿画像も素性として利用することで正解率を向上させた．具体的には，画像とテキストのそれぞれを素性とする分類器を別々に学習し，その結果を素性としたアンサンブルモデルを学習した．実験の結果，画像を素性とする分類器とのアンサンブルにより，従来のテキストのみの分類器と比べて，性別推定の性能が向上したことを確認した．

Ma らは，ユーザが投稿したツイートの画像を手掛かりに Twitter ユーザの性別を予測する手法を提案した [9]．画像に対し，その被写体の種類を表す 10 種類のラベルと，それを投稿したユーザの性別のラベルを付与したデータセットを構築した．このデータセットを用いて，画像の被写体を分類するモデル，並びに被写体から性別を予測するモデルを学習した．評価実験により，ツイートの画像がユーザの性別推定に有効であることを示した．

Liu らは，Twitter における投稿テキスト，自己紹介文，Twitter に関する統計量を素性とし，機械学習の古典的モデルならびに深層学習モデルについて，Twitter ユーザのプロフィール推定の性能を比較した [8]．推定するプロフィールは性別を含む．学習モデルとして，Word Embeddings，Sequence Embeddings，ロジスティック回帰，サポートベクトルマシーン，ナイーブベース，決定木モデルを比較したり，古典的モデルの素性として Unigram，Bigram，Sequential Pattern を比較したりした．最終的に Bidirectional Encoder Representations from Transformers(BERT)[5] を用いて単語埋め込みを獲得し，ツイートに含まれる単語の埋め込みの平均ベク

トルと，自己紹介文に含まれる単語の埋め込みの平均ベクトルを作成し，これを特徴ベクトルとして全結合層への入力としたニューラルネットワークモデルが一番良い結果を出した．また，Twitter における多様な情報を考慮することの有効性を示した．

Wang らはプロフィール画像，ユーザ名，ユーザスクリーン名，自己紹介文をもとに性別，年齢，所属組織を推定する分類器を学習した [15]．ツイートをもとに Twitter ユーザのプロフィールを推定する関連研究とは異なり，これを用いずに他の情報によってプロフィールを推定することの有効性を示した．また，使用した素性のうちどれか 1 つを使った時でも十分に高い正解率が得られることを確認した．

2.1.2 性別以外のプロフィールの予測に関する研究

Morgan-Lopez らは，誕生日ツイートなどの Twitter 特有の特徴量を利用することで，ユーザの年齢推定の精度を向上させた [10]．Liu らの論文 [8] で用いられた Twitter 統計量に加え，独自の素性を提案・使用することで，十分に高い推定精度が得られることを示した．

Seol らは，Twitter におけるフォロワー間のネットワーク関係を学習し，政治家のフォロワーを支持者，中立者，非支持者の 3 つに分類するモデルを作成した [12]．具体的には，あるユーザが政治家に対して言及したツイート率やリツイート率を測定し，またその言及したツイート，リツイートの内容が肯定的であるか否定的であるかを分類し，そのユーザによる複数のツイートの極性を総合的に分析して，政治家に対する立場を分類した．

杉谷らは，位置情報が付与されていないツイートに対し，そのツイートに書かれている地名やスポット名からツイートの投稿位置を推定する手法を提案した [13]．データセットとして，位置情報付きツイートに対して，スポット名とそのスポットの緯度，経度を付与したものを用意した．ツイートにスポット名が出現したときに，そのスポット位置が実際にツイートを投稿した位置と一致するかを判断する分類器を学習し，これによって一致すると判定された場合に，ツイートの位置を特定した．実験では，実際のツイートの位置と推測した位置の誤差が 1km 以内の場合を正解，それ以外を不正解として評価した．実験の結果，提案手法によって精度 80%，再現率 43%でツイートの位置を特定できた．「みぞれ」や「リアル」など一般的な単語がスポット名になっているときに地名でない単語を地名と誤って判断した事例や，スポット名の位置として登録されている緯度，経度がずれていたことによって位置の推定に失敗した事例があった．この研究は，厳密にはユーザのプロフィールの推定ではないが，ツイートからユーザの情報 (この場合は位置) を推定しているという点で本研究と関連している．

2.2　言語モデル・機械学習モデル

提案手法では，いくつかの既存の言語モデルや機械学習モデルを利用する．本節ではこれらのモデルの概要を紹介する．

2.2.1　BERT

BERT[5] は大量のテキストから学習された汎用言語モデルであり，Transformer 層と呼ばれるユニットから構成されるニューラルネットワークモデルである．Transfomer 層は，ベクトルの時系列データが与えられたときに，その系列の前後に出現するベクトルとの関連性を考慮し，時系列におけるある地点の埋め込み表現を獲得する．BERT の場合，テキストは単語列として表現され，単語を表すベクトルの時系列が入力となる．Transformer 層は，図 2.1 に示すように，Multi-head Attention 層，正規化層 (Add&Norm)，Feed-forward Network 層，正規化層，がこの順番で重ねられている．

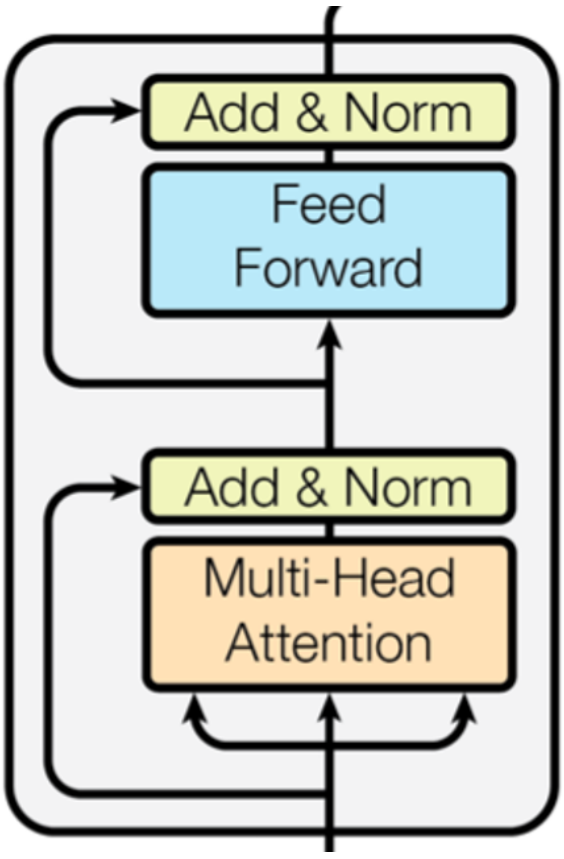

図 2.1: Transformer 層

Multi-head Attention 層は，式 (2.1) に示すように，入力されたベクトル $\mathbf{X}$ に対し，重み $\mathbf{W}^q$，$\mathbf{W}^k$，$\mathbf{W}^v$ との外積によって，$\mathbf{Q}$(クエリ)，$\mathbf{K}$(キー)，$\mathbf{V}$(バリュー)

を得る．得られた $\mathbf{Q}$，$\mathbf{K}$，$\mathbf{V}$ を用いて，式 (2.2)，(2.3) によって $\mathbf{A}$ を計算する．Multi-Attention 層では，直感的にはある時点の埋め込みベクトルを得る際に，他のどの時点の要素を重要視するかを学習する．

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^q,\ \mathbf{K} = \mathbf{X}\mathbf{W}^k,\ \mathbf{V} = \mathbf{X}\mathbf{W}^v \tag{2.1}$$

$$Softmax(x_i, \mathbf{x}) = \frac{e^{x_i}}{\sigma_{k=1}^{n} e^{x_i}} (i = 1, 2, ...n) \tag{2.2}$$

$$\mathbf{A} = Softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})V \tag{2.3}$$

正規化層では，式 (2.4) に示すように，入力ベクトル $\mathbf{y}_i$ が与えられたとき，その要素の平均 μ_i と標準偏差 σ_i を用いて正規化する．ここで β と γ はパラメータであり，$\mathbf{y}_i$ と同じ次元のベクトルである．また $\bigodot$ はベクトルの要素ごとに積をとる演算を表す．

$$LayerNorm(y_i) = \frac{\gamma}{\sigma_i} \bigodot (\mathbf{y}_i - \mu_i) + \beta \tag{2.4}$$

Feed-forward Network 層は 2 層のフィードフォワードネットワークで構成される．入力を $\mathbf{z}_i$，重み行列を $\mathbf{W}_1$，$\mathbf{W}_2$，バイアスを $\mathbf{b}_1$，$\mathbf{b}_2$ とし，$\mathbf{z}_i$ に対するフィードフォワード層の出力 $FFN(Z_1)$ は，式 (2.5)，(2.6) のように計算される．

$$FFN(\mathbf{z}_i) = GELU(\mathbf{z}_i W_1 + \mathbf{b}_1)W_2 + \mathbf{b}_2 \tag{2.5}$$

$$GELU(x) = x \cdot \frac{1}{2}[a + erf(\frac{x}{\sqrt{2}})] \tag{2.6}$$

BERT では，上記の 3 つの要素で構成される Transformer 層を 12 層重ねることで，文の深い意味表現が得られるように設計されている．

BERT の学習は，事前学習とファインチューニングという 2 つの手続きから構成される．事前学習では，大量のテキストから文の一般的な抽象表現 (意味表現) をあらかじめ学習する．事前学習は以下の 2 つのタスクを解くことで実現される．1 つ目は，ランダムに選択した単語を [MASK] というトークンに置き換え，その [MASK] に埋めるべき単語を予測するといった穴埋め問題を解くタスク (Masked Language Model Task) である．2 つ目は，文の組を与え，1 つの文がもう 1 つ文に続いて出現するか否かを判定するタスク (Next Sentence Prediction Task) である．次に，ファインチューニングでは，自然言語処理のある特定の問題 (タスク) を解くために，そのタスクのラベル付きデータを用意し，事前学習した BERT モデルのパラメタをそのタスクに合わせて更新する．このとき，ラベル付きデータの量は比較的少量でよいとされている．すなわち，事前学習で汎用的なモデルを学習

し，これをタスクに合わせて調整することで，優れた性能を持つモデルを学習する．BERT は，文分類，極性判定，含意関係認識など，自然言語処理における様々なタスクに応用でき，高い正解率が得られることが報告されている．本研究が目的とする Twitter ユーザの性別推定にも応用することが可能である．

2.2.2 Vision Transformer

Vision Transformer[4] は，与えられた画像を特徴ベクトルに変換するモデルの 1 つである．BERT 同様に，12 層の Transformer 層で構成されたモデルである．Transformer 層は入力が時系列であることを想定しているのに対し, 画像は 2 次元のデータである．そのため，まず画像を格子状に分割する．分割した画像を左上から右下の順に並べ，1 次元の時系列の情報に変換する．この情報に，時系列におけるそれぞれの時点が画像全体のどの位置に属しているかの情報を加えたものが Vision Transformer の入力となる．そして，大量の画像データから，画像の一般的な特徴ベクトルを出力するモデルを事前学習する．

2.2.3 Light GBM

Light Gradient Boosting Machine(Light GBM) は，決定木をベースとした機械学習アルゴリズムもしくは分類器である．Gradiend Boosting Machine とは，損失関数を下げることを目的として弱い学習器を複数集め，強い学習器を作成する方法である [7]．Light GBM では弱い学習器として決定木を用いる．決定木とは，ある特徴量に対して「はい」か「いいえ」で答えられる質問を用意し，質問を繰り返すことで入力データを分類するモデルである．質問の答えが「はい」か「いいえ」によってデータが分岐され，これが階層的に繰り返されるため，モデル全体は木構造で表現される．一般の決定木モデルと Light GBM の違いとして以下の点が挙げられる．

Leaf-wise tree growth 一般的な決定木は分類されるまでに必要な質問回数が均等になるように学習されるが，Light GBM ではそのような制約は考慮されず，質問の必要がないときはその時点でデータの分類ラベルが決まる．

Histogram based 質問を学習する際には質問の良さを評価するが，いくつかのデータをまとめてヒストグラムを作成し，このヒストグラムを元に質問の評価値を計算することで，その計量量を削減する．

Gradient-based one-side sampling 学習できていない要素を優先するために，誤差が小さいデータは減らし，誤差の大きいデータだけを残すことで訓練データ全体の量を減らす．

Exclusive feature bundling 学習に用いる素性のうち，まとめても学習に影響のない複数の素性はひとつにまとめ，素性数を削減することで計算量を減らす.

2.3 本研究の特色

Twitter ユーザを対象とした性別推定の先行研究の多くは，ツイート (テキスト) を素性とした分類器を学習する．Sasaki らの研究 [11] のように，ツイートのテキストと画像を考慮した手法もあるが，テキストと画像で異なる分類器を学習してから，これらを組み合わせている．これに対し, 本研究では，ツイートのテキストと画像を同時に考慮した分類器を設計・学習する事で，テキストと画像の関連性を捉え，これにより性別推定の正解率を向上させる手法を探求する．一方，ツイートだけではなく，プロフィール画面など，Twitter の様々な情報を素性として性別を予測する研究も行われている [8]．本研究では，より多角的な情報をもとにユーザの性別を正確に予測することを狙う．すなわち，どの先行研究よりも多くの種類の情報を同時に考慮したモデルを提案する．

第3章　提案手法

本章は，本研究の提案手法について述べる．本研究は，図 3.1 に示す 4 つの手続きによって進められる．始めに，[ユーザの収集] では，性別の自動ラベル付けが可能な Twitter ユーザを選別し，収集する．次に，[Twitter 情報の収集] では，取得したユーザのプロフィール情報やツイート情報を収集する．次に，[分類器の学習] では，収集した情報をもとに複数の分類器を学習する．最後に，[データ拡張] では，学習した分類器をもとにラベル付きデータを自動的に構築し，元の訓練データに加えることで訓練データを拡張する．また，拡張後の訓練データを用いて性別推定の分類器を再学習する．

以下，これらの手続きの詳細を説明する．3.1 節では，[ユーザの収集] と [Twitter 情報の収集] について述べる．3.2 節では，[分類器の学習] について述べる．最後に 3.3 節では，[データ拡張] について述べる．

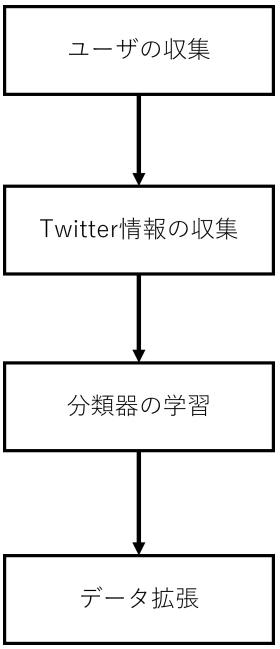

図 3.1: 実験の流れ

3.1 データセットの構築

ユーザの性別を予測するモデルを学習するための訓練データとして，Twitter のユーザ，その性別ラベル，そのユーザに関する情報 (ツイートやプロフィール画像など) を紐づけたデータセットを構築する．まず，Twitter ユーザのうち，著名人の性別は自動付与しやすいと考え，著名人の Twitter アカウントを収集する．次に，著名人の Wikipedia の情報などから性別ラベルを自動的に付与する．その後，性別ラベルを付与したユーザについて，そのユーザが投稿したツイートや画像などの情報を取得する．

3.1.1 対象ユーザの選定

既に述べたように，データセットに収録するユーザは著名人とする．著名人はプロフィールを公開していることが多く，また Wikipedia といった外部情報に性別を推測できる情報が記載されていることもあり，性別を自動的にラベル付けできる可能性が高いと考えたためである．具体的には，Twitter 日本フォロワー数総合ランキング [14] から，上位 2 万件のアカウントを収集する．フォロワー数が多いユーザの多くは，著名人であることを確認している．

3.1.2 性別の自動ラベル付け

収集したユーザに対して，性別の自動ラベル付けを試みる．その手続きを以下に述べる．

1. Twitter ユーザの Wikipedia のページを取得する．図 3.2 に示すように，まず対象ユーザの「ユーザ名 Wiki」を検索キーワードとし，検索エンジン Google を用いてウェブ検索を行う．検索結果の上位 5 件以内に，Wikipedia へのリンクがあるとき，その Wikipedia のページの URL を取得する．見つからない場合，そのユーザはデータセットから除外する．

2. 人物の Wikipedia のページには，Infobox にその人物のプロフィールが記載されているが，性別の情報は記載されていない．そこで，本人の画像から性別を自動推定する．具体的には，図 3.3 の左に示すように，取得した Wikipedia に掲載されているユーザのプロフィール画像を取得する．次に，その画像に対し，Face++[6] を用いて人物の性別を推定する．Face++はオンライン顔認識プラットフォームであり，画像中の顔を認識することができる．さらに，認識した人物の性別を推定することもできる．Face++は高い精度で顔認識ならびに性別推定ができることが知られている．そのため，本研究では Face++による性別判定結果を Twitter ユーザの性別ラベルとして付与する．

3. Wikipedia のページに人物のプロフィール画像が掲載されていないときは，画像検索により，その人物の顔の画像を収集する．具体的には，図 3.3 の右上に示すように，Wikipedia のエントリ名 (人物名) を検索キーワードとして Google 画像検索を行い，その検索結果の上位 5 件の画像を取得する．Face++ を用いて各画像の性別を判定し，その多数決によって性別ラベルを付与する．

図 3.3 の右下の枠内は,「松本人志」について,「Man:1」は Wikipedia のプロフィール画像を Face++で判定した結果が男性であること,「Man:5」は Google 画像検索の上位 5 件の画像を Face++で判定した結果の全てが男性であることを示している．

図 3.2: ユーザの Wikipedia ページの取得

3.1.3 Twitter 情報の収集

Twitter ユーザを選定し，その性別ラベルを決定した後，そのユーザの性別を自動的に推定するための情報 (以下，Twitter 情報と呼ぶ) を Twitter から取得する．

3.1.3.1 収集する情報

Twitter には，様々な情報が記載されているが，本研究では図 3.4 に示す 6 種類の情報を収集する．それぞれの情報の詳細を以下に示す．

1. **ヘッダー画像**: Twitter のユーザページの上部に表示される横長の画像．この画像はユーザが自由に設定できるが，内容や色調など，男女によって掲載する画像に違いがあると考えられる．

図 3.3: 人物の画像の取得

2. **プロフィール画像**: ユーザのアイコンとなる画像．ツイート時やフォロー一覧などに表示される．プロフィール画像にユーザの顔がある場合，性別判定の有力な手掛かりになると考えられる．

3. **ユーザ名**: ユーザの名前．プロフィール画像と同様にツイート時やフォロー一覧などに表示される．ユーザ名としてニックネームが使われることもあれば，本人の名前が使われていることもある．男性らしい名前，女性らしい名前があるように，本人の名前は性別推定の有力な手掛かりになる．

4. **自己紹介文**: ユーザ自身によって書かれた自己紹介文．ユーザページのプロフィール欄のみに表示される．自分の紹介をしている文章であるため，その内容や書き方に性差が表れると考えられる．

5. **フォロー情報**: そのユーザがフォローしている別のユーザ (フォロイー) の情報を取得する．

6. **ツイート情報**: ユーザが投稿したツイートを取得する．テキストだけでなく投稿画像も取得する．

3.1.3.2 情報収集の詳細

情報の収集には TwitterAPI を使用する．プロフィール画像，ヘッダー画像については，画像の URL が取得できる．その URL から画像をダウンロードする．ユーザ名，自己紹介文はユーザページに表示されているテキスト情報のみを収集する．

Twitter API では，Twitter のユーザページ上で「フォロワー」と表示されるユーザ (そのユーザをフォローしている別のユーザ) の人数や「フォロー中」と表示され

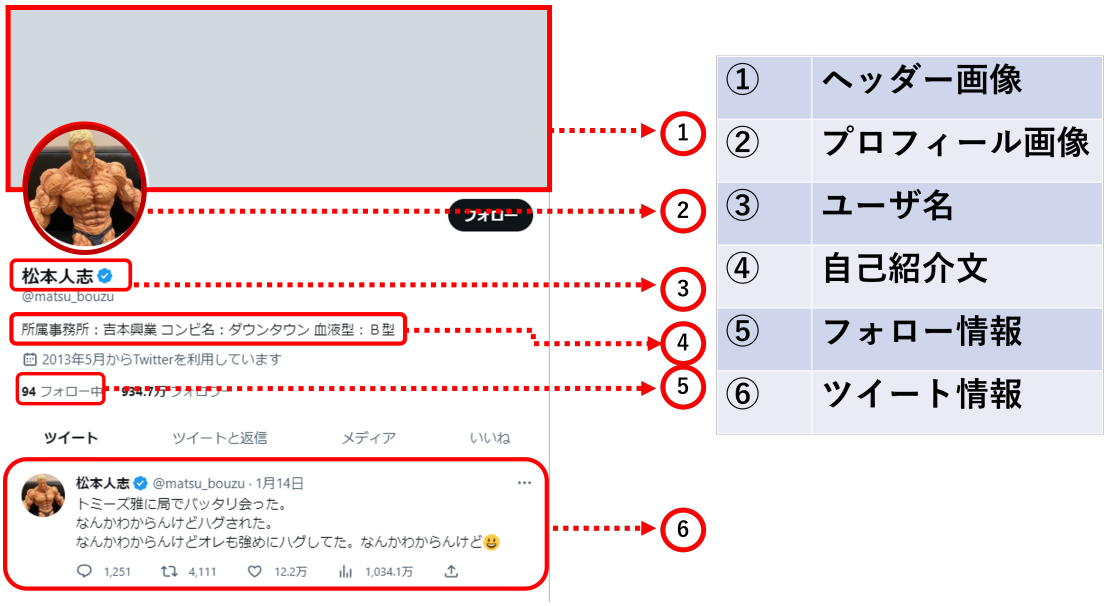

図 3.4: Twitter のユーザページの例

るユーザ (そのユーザがフォローしている別のユーザ，すなわちフォロイー) の人数だけでなく,「フォロー中」のユーザ名も取得できる．そこで対象ユーザのフォロイーを最大 100 件ランダムに選び，そのアカウント情報を取得する．さらに，フォロイーの自己紹介文を収集する．

ツイート情報として，ユーザが投稿したツイートを取得する．画像と一緒にツイートを投稿している場合，テキストと画像の両方を取得する．ツイートには自身で書き込むツイートのほかに，リツイートと呼ばれる他のユーザのツイートを参照したツイートがある．ここではリツイートは収集の対象とせず，自身のツイートを最大 100 件収集する．より具体的には，ツイートテキスト，ツイート画像，ツイートした日時を取得する．

この他に，Twitter API では現在地や誕生日などの情報も取得できるが，本研究では使用しないため，収集しない．

3.2 分類器の学習

3.2.1 概要

本研究では，性別推定タスクを以下のように定義する．与えられた Twitter ユーザに対し，その性別を表すスコアを予測する．性別スコアは，男性の時は 1，女性の時は 0 とする．また，1 に近い値ほど男性である可能性が高く，0 に近い値ほど女性である可能性が高いと解釈する．以下，性別スコアを出力するモデルを単に「分類器」と呼ぶ．

本研究では，性別予測のために複数の情報を用い，それぞれの情報を入力とした分類器を複数学習する．本研究で学習する分類器を以下にまとめる．

- ツイートを対象とした分類器．ツイートのテキストと画像の情報を素性として用いる．
- Twitter 統計量による分類器
- ユーザ名による分類器
- プロフィール画像，ヘッダー画像による分類器
- 自己紹介文による分類器
- フォロイーの自己紹介文による分類器
- 統合モデル．上記の分類器を組み合わせた分類器．

3.2.2 項以降では，これらの分類器の詳細を説明する．
上記の分類器を実装するために用いるモデルについて述べる．

- 2.2.1 項で紹介した BERT は，テキストを入力とした分類器を学習するために用いる．具体的には，ツイートテキストや自己紹介文を用いた分類器の学習に用いる．
- 2.2.2 項で紹介した Vision Transformer は，画像を入力とした分類器を学習するために用いる．具体的には，ツイート画像，プロフィール画像，ヘッダー画像を用いた分類器の学習に用いる．
- 2.2.3 項で紹介した Light GBM は，主にテキスト・画像以外の情報を入力とした分類器の学習に用いる．具体的には，Twitter 統計量，ユーザ名，フォロイーの自己紹介文を用いた分類器の学習に用いる．また，統合モデルの学習にも用いる．

3.2.2 ツイートを対象とした分類器

ここではユーザによって投稿された複数のツイートからそのユーザの性別を予測する．特に，ツイートのテキストと画像の両方を考慮し，これらの組み合せ方によって複数のモデルを提案する．
性別推定は，(1) 個々のツイートに対する性別推定，(2) ユーザの性別推定，の 2 段階で行う．詳細を以下に示す．

1. 個々のツイートに対し性別スコアを予測する．既に述べたように，性別スコアは，1 のときは男性，0 のときは女性を表すものとする．

2. ユーザが投稿した複数のツイートの性別スコアからユーザの性別を決定する．この際，以下の 3 通りの方法を採用する．

   Tw-ave ツイートの性別スコアの平均を求め，0.5 以上のとき男性，それ未満のとき女性と判定する．以下，この方法を Tw-ave と記す．

   Tw-soft ツイートの性別スコアを Softmax 関数を用いて正規化した後，先ほどと同様に性別スコアの平均を求め，0.5 を閾値として性別を決定する．以下，この方法を Tw-soft と記す．

   Tw-sel Softmax 関数による正規化の後，性別クラス毎に性別スコアが 1 または 0 に近い上位 10%のツイート (合計 20%) を選別した後，先ほどと同様に性別スコアの平均を求め，0.5 を閾値として性別を決定する．以下，この方法を Tw-sel と記す．

以下，ツイートを対象にした個々の分類器の詳細について述べる．なお，Text only model と Image only model を除いて，ツイートのテキストと画像を同時に考慮した深層学習モデルとなっている．

3.2.2.1 前処理

分類器を学習する前に，ツイートのテキストに対して以下の前処理を行う．

- **テキスト以外の文字列の削除**: URL，リプライ (他のユーザへのツイートの返答) で表示される相手ユーザ名，ハッシュタグを削除する．
- **絵文字の削除**: 絵文字を削除する．後述する形態素解析ツール MeCab は絵文字に対応していないため，事前に絵文字を削除することで形態素解析の誤りを減らす．
- **形態素解析**: BERT へは単語の系列を入力として与える必要があるため，形態素解析ツール MeCab を用いてツイートのテキストを単語に分割する．

一方，ツイートの画像に対しては以下の前処理を施す．

- **画像サイズの統一**: ツイートに添付される画像のサイズは統一されていないので，Python の画像編集ライブラリ Pillow[2] を用いて 224×224 のサイズに統一する．1 つの点の色は RGB(赤，緑，青) の 3 値で表現されるため，画像データの大きさは $3 \times 224 \times 224$ となる．

3.2.2.2 Text only model

ツイートのテキストのみを用いて，性別を推定する分類器である．そのアーキテクチャを図3.6に示す．まず，訓練データを用いてBERTをfine-tuningする．前処理済みのツイートテキストをBERTに入力し，768次元の特徴ベクトルを得る．次に，BERTから得られた特徴ベクトルを全結合層 (Fully Connected Layer;FCL) に渡す．その詳細を図3.5に示す．最後に全結合層の出力をSoftmax関数にかけ，0以上1以下の性別スコア (Gender score) を出力する．

本研究はテキストと画像を同時に考慮した性別推定モデルの学習に取り組むが，Text only modelはテキストのみを用いるモデルであり，提案手法との比較に用いる．

- 入力は前の層の特徴ベクトル，出力は性別クラスの2次元ベクトルである．
- 入力ノードと出力ノードの全ての組にリンクを張る．(全結合)
- 活性化関数として$tanh(\mathbf{x})$を使用する．
- モデル学習時の設定
  - 学習時の損失関数はLogLossとする．y_iとd_iはi番目のデータに対する予測値と正解値で，nは訓練データの総数である．

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n} d_i \log y_i + (1-d_i)\log(1-y_i)$$

  - Dropout率は0.3とする．

図 3.5: ツイートに対する分類器における全結合層

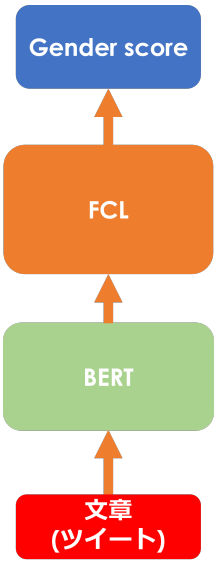

図 3.6: Text only model

3.2.2.3 Image only model

ツイートに投稿された画像のみを用いて，性別を推定する分類器である．そのアーキテクチャを図 3.7 に示す．具体的には，まず，訓練データによって fine-tuning した Vision Transformer によって 384 次元の画像の特徴ベクトルを得る．次に，得られた特徴ベクトルを図 3.5 に示す全結合層に渡し，Softmax 関数をかけて性別スコアを得る．なお，ツイートに画像が含まれていないときは，モデルの学習の際には訓練データから除外する．また，予測時には，画像のないツイートは判定不能とみなす．Text only model と同様に，Image only model は，テキストと画像を同時に考慮する提案手法との比較に用いる．

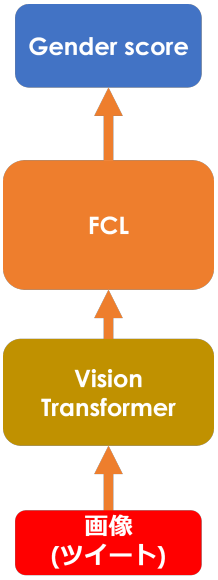

図 3.7: Image only model

3.2.2.4 Early fusion model

ツイートのテキストと画像を同時に考慮して，性別を判定する．そのアーキテクチャを図 3.8 に示す．テキストは BERT，画像は Vision Transformer によってそれぞれの 768,384 次元の特徴ベクトルに変換する．ただし，画像がないツイートに対する画像の特徴ベクトルはゼロベクトルとする．次に，これらを連結した 1152 次元のベクトルを図 3.5 に示す全結合層に渡し，Softmax 関数をかけて性別スコアを得る．BERT，Vision Transformer から出力された特徴量をそのまま結合することで，両方の特徴を考慮した性別推定モデルが学習できると考えられる．

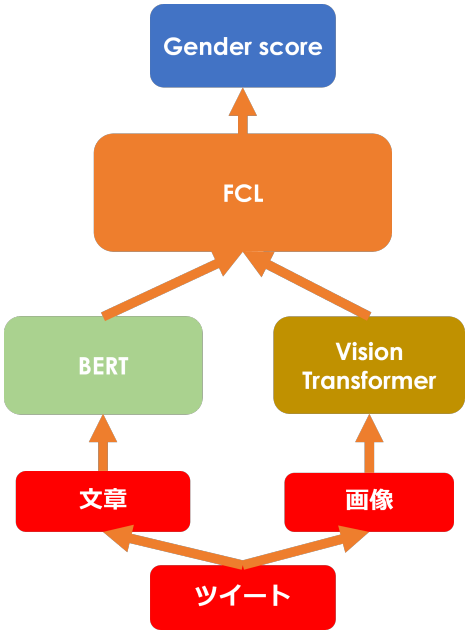

図 3.8: Early fusion model

3.2.2.5 Late fusion model

Early fusion model と同じく，ツイートのテキストと画像を同時に考慮するが，両者の特徴ベクトルを次元圧縮した後で組み合わせる．アーキテクチャを図 3.9 に示す．テキストと画像をそれぞれ BERT と Vision Transformer によって 768，384 次元の特徴ベクトルに変換する．これらの特徴ベクトルを，それぞれ図 3.5 に示す全結合層 (ただし，出力のノード数は 4) に渡し，4 次元のベクトルに圧縮する．これらを連結した 8 次元のベクトルを，図 3.5 に示す別の全結合層に渡し，Softmax 関数をかけて性別スコアを得る．Early fusion model との違いは，テキストと画像の特徴ベクトルを一度個別に次元圧縮した後，組み合わせる点である．どちらの手法がテキストと画像を同時に考慮するモデルとして適切か，実験により検証する．

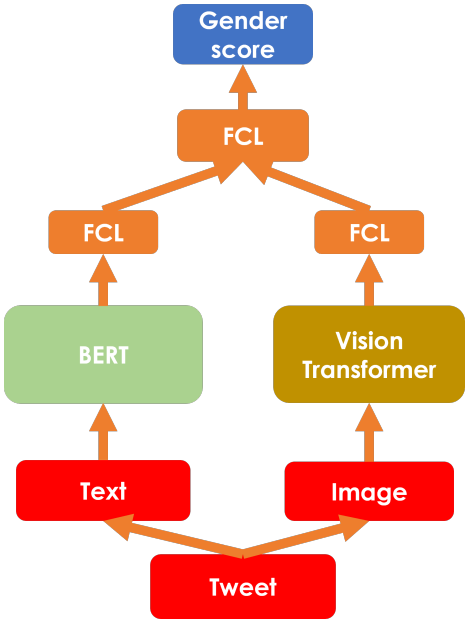

図 3.9: Late fusion model

3.2.2.6 Dense fusion model

Early fusion model と Late fusion model を組み合わせたモデルである．そのアーキテクチャを図 3.10 に示す．BERT によってテキストを 768 次元の特徴ベクトルに，Vision Transformer によって画像を 384 次元の特徴ベクトルに，それぞれ変換する．次に，Early fusion model と同じ方法で全結合層 (ただし，出力のノード数は 8) に渡し，8 次元の特徴ベクトルを得る．それと同時に，Late fusion model と同じ方法で中間層の 8 次元の特徴ベクトルを得る．これらのベクトルを連結して，16 次元の特徴ベクトルを作成し，これを図 3.5 に示す全結合層に渡し，Softmax 関数をかけて性別スコアを得る．モデル全体の構造が大きくなり，計算コスト，メモリ使用量が多くなるが，Early fusion model と Late fusion model の双方の利点が生かされ，性別推定の性能の向上が期待できる．

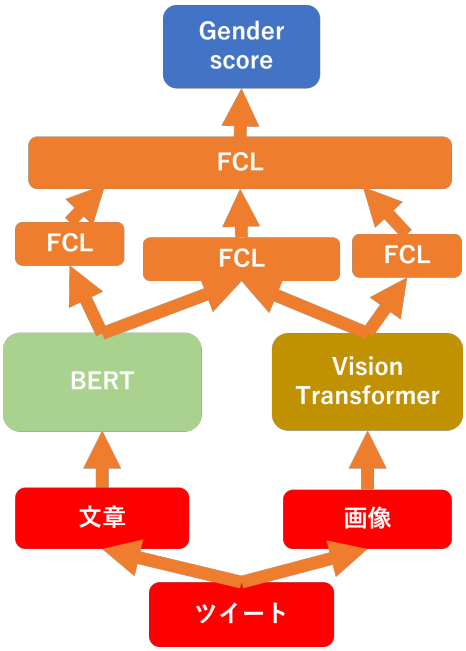

図 3.10: Dense fusion model

3.2.2.7 Caption model

ツイートのテキストと画像を同時に考慮するが，画像をキャプション (テキスト) に変換してから両者を組み合わせる．Caption model のアーキテクチャを図 3.11 に示す．まず，Clip model[3] を用いて画像の英語のキャプションを生成する．次に，Google translate[1] によって英語のキャプションを日本語に翻訳する．そして，ツイートのテキストと日本語のキャプションを特殊トークン ⟨SEP⟩ を挟んで連結する．これを入力テキストとして，BERT を用いて 768 次元の特徴ベクトルを得る．これを図 3.5 に示す全結合層に渡し，Softmax 関数をかけて性別スコアを得る．キャプション生成によって，画像の中でも被写体など重要な部分をテキストとして表現し，これを素性として用いることで，画像全体の特徴ベクトルを用いるモデルよりも，性別推定に有効な情報が効率的にモデルに反映されることを期待している．

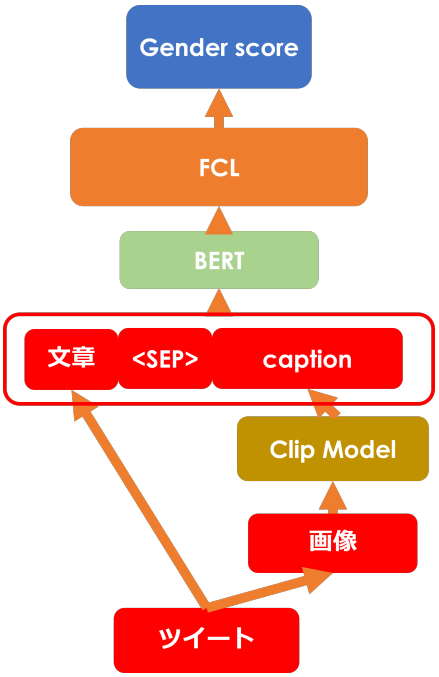

図 3.11: Caption model

3.2.2.8 Ensemble of Text only model and Early fusion model

Text only model と Early fusion model を組み合わせたモデルである．そのアーキテクチャを図 3.12 に示す．画像のないツイートは Text only model で，画像付きのツイートは Early fusion model で性別スコアを出力する．予備実験の結果，画像を含まないツイートに対しては，テキストと画像を同時に考慮するモデルよりも，テキストのみを考慮するモデル (Text only model) の正解率が高いことが分かった．一方，画像を含むツイートに対しては，Early fusion model の正解率が最も高かった．このため，画像の有無によって両者を使い分けることで，性別推定の正解率を向上させることを狙う．

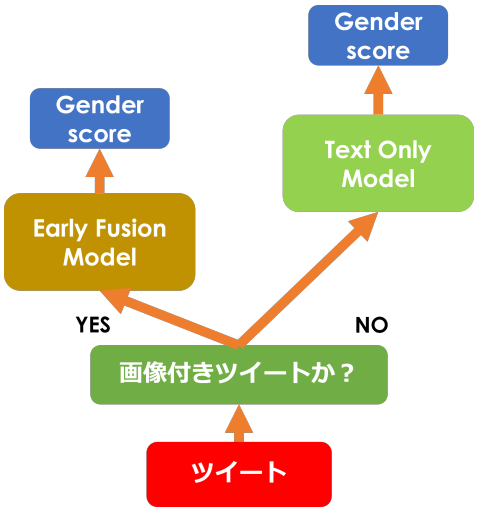

図 3.12: Ensemble of Text only model and Early fusion model

3.2.3 Twitter 統計量による分類器

Liu らの研究 [8] を参考に，Twitter から得られる統計情報を素性として分類器を学習する．使用した Twitter 統計量の一覧を表 3.1 に示す．学習アルゴリズムとして Light GBM を用いる．

表 3.1: 性別推定に使用する Twitter 統計量

一日の平均ツイート数，一時間ごとのツイート率，曜日ごとのツイート率
ツイートの平均統計量 (文字数，ハッシュタグ数，単語数，特殊文字数，句読点数，漢字数)
自己紹介文の統計量 (文字数，ハッシュタグ数，単語数，特殊文字数，句読点数，漢字数)
フォロー数，フォロワー数

3.2.4 ユーザ名による分類器

Twitter ユーザの名前から性別を予測する．具体的には，ユーザ名の文字の uni-gram，ひらがな表記のユーザ名の文字 bi-gram を素性とする．また，「〜夫」「〜子」などユーザ名の末尾に性別を表す文字が出現することが多いことから，末尾の文字 uni-gram，ひらがな bi-gram に ‘end_’ という特殊記号をつけたものも素性とする．ユーザー名から得られる素性の例を表 3.2 に示す．学習アルゴリズムとして Light GBM を用いる．

表 3.2: ユーザ名の素性の例

ユーザ表示名：松本人志
読み：まつもとひとし
分割例：’まつ’，’つも’，’もと’，’とひ’，’ひと’，’とし’，’end_とし’，’end_し’，’松’，’本’，’人’，’志’，’end_志’

3.2.5 プロフィール画像，ヘッダー画像による分類器

ユーザのプロフィール画像とヘッダー画像から性別を判定する．プロフィール画像のみ，ヘッダー画像のみ，両方を素性とする 3 つの分類器を学習する．分類モデルとして Vision Transformer を用いる．以下，分類器の詳細について述べる．

プロフィール画像のみとヘッダー画像のみを用いる分類器では，まず，Vision Transformer によってそれぞれの画像を 384 次元の画像特徴ベクトルに変換する．次に，得られた特徴ベクトルを全結合層に渡し，Softmax 関数にかけて性別スコアを得る．

プロフィール画像とヘッダー画像の両方を素性とする分類器では，Vision Transformer によって 384 次元の特徴ベクトルに変換する．次に，これらを連結とした 768 次元のベクトルを全結合層に渡し，Softmax 関数をかけて性別スコアを得る．

プロフィール画像とヘッダー画像の片方もしくは両方の画像がない場合，そのユーザは学習に用いない．また予測の際にはそのユーザに対する性別は予測不能とする．

例として，図 3.13 に「松本人志」のプロフィール画像とヘッダー画像を示す．プロフィール画面はユーザ本人の全身写真であり，ボティービルダーのような体付きをしていることから，ユーザが男性であることを示唆していると言える．この例からわかるように，プロフィール画像は性別判定の有力な手がかりになりうる．一方，ヘッダー画像は存在しない．そのため，このユーザはプロフィール画像のみを用いる分類器を学習する際に利用するが，ヘッダー画像のみを用いるモデル，両方を用いるモデルの時には学習に利用しない．また，ヘッダー画像のみを用いるモデル，両方を用いるモデルは，このユーザの性別は予測不能とする．

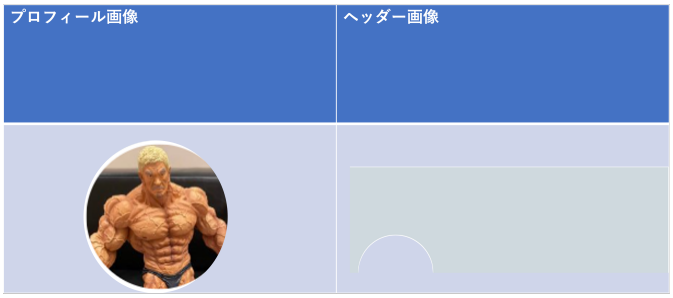

図 3.13: プロフィール画像とヘッダー画像の例

3.2.6　自己紹介文による分類器

Twitter 上のプロフィール画面に表示される自己紹介文から性別を予測する．すなわち，自己紹介文のテキストを素性として，性別を推定する分類器を学習する．分類モデルとして BERT を用いる．一般に，自己紹介は，表 3.3 の例のように箇条書きや複数の文で書かれることもある．一方，BERT の入力は 1 つの文である．そこで，自己紹介文の文字列を単に 1 つの文として扱い，BERT に入力し，性別を予測する．

表 3.3: 自己紹介文の例

松本人志	所属事務所：吉本興業 コンビ名：ダウンタウン 血液型：B 型

3.2.7　フォロイーの自己紹介文による分類器

Twitter のユーザの中には，情報を閲覧することを主な目的として Twitter を活用する人もいる．そのようなユーザは，ツイートの投稿件数が少なく，またプロフィール画像や自己紹介文といった自身に関する情報を開示していないことが多い．これまでに述べてきた分類器は，これらの情報を手がかりにしているため，情報発信を積極的に行わないユーザに対する性別を正確に推定できない可能性がある．

ここで，ユーザがフォローしている別のユーザ，すわわちユーザのフォロイーに着目する．フォロイーは，ユーザが関心を持っている人々であり，ある程度性差があると考えられる．特に，著名人がフォロイーになるとき，女優や女性アイドルなど男性にフォローされやすい人と，女性にフォローされやすい人がいる．すなわち，フォロイーは性別を推定する手がかりとなりうる．また，自身では情報を積極的に発信しないユーザに対し，フォロイーの情報を使うことで性別推定のための素性を補完する効果も期待できる．

そこで，性別推定の対象となるユーザの複数のフォロイーについて，その自己紹介文を用いて対象ユーザの性別を予測する．対象ユーザのフォロイーをランダムに 100 名選択し，その自己紹介文を取得する．これらの自己紹介文を MeCab を用いて形態素解析し，単語を素性，その TF-IDF を値とする素性ベクトルを作成する．さらに，Latest Semantic Indexing(LSI) によって，素性ベクトルを 1000 次元に圧縮する．これを素性として，性別推定の分類器を Light GBM を用いて学習する．

3.2.8　統合モデル

これまで述べてきた分類器を統合した性別推定モデルを学習する．3.2.2 項で述べたツイートを用いる分類器のうち 1 つと，3.2.3 項の Twitter 統計量による分類

器，3.2.4 項のユーザ名による分類器，3.2.5 項のプロフィール画像による分類器，ヘッダー画像による分類器，プロフィール画像とヘッダー画像の両方による分類器，3.2.6 項の自己紹介文による分類器，3.2.7 項のフォロイーの自己紹介文による分類器を組み合わせる．そのアーキテクチャを図 3.14 に示す．具体的には，各分類器が出力する性別スコアを素性とし，Light GBM を用いて最終的な性別スコアを得る．ただし，ヘッダー画像のないユーザについてはヘッダー画像による分類器では性別を判定できないなど，入力として必要な情報がない場合には個々の分類器によって性別スコアを推定できないことがある．このような場合，その分類器が出力する性別スコアは 0.5 として，統合モデルの素性としている．

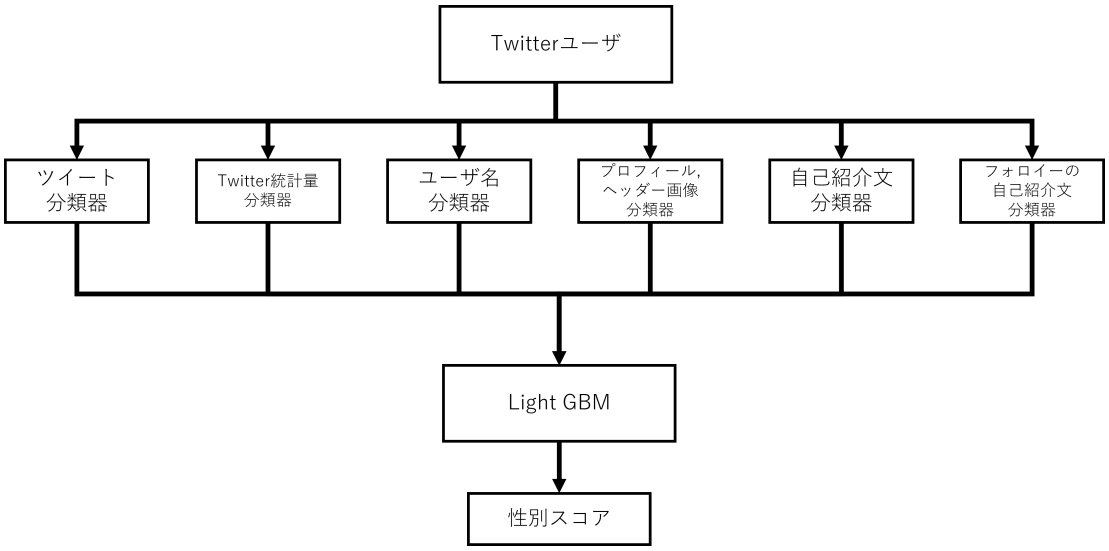

図 3.14: 統合モデル

3.3 データ拡張

Twitter ユーザの性別を推定する分類器を学習するためには，性別のラベルが付与されたユーザのデータが必要である．人手によるラベル付けはコストが高いため，大規模な性別のラベル付きデータを用意することは難しい．そこで，ラベル付きデータを自動的に生成し，訓練データを拡張する．

データ拡張のもう一つの目的は，著名人ではなく一般の Twitter ユーザの性別推定の性能を向上させることにある．3.1 節で述べたように，本研究の訓練データは，主に著名人のユーザを収集し，これに自動的に性別のラベルを付与したものである．そのため，学習されたモデルは，著名人の性別は正確に推定できるが，一般ユーザの性別推定の正解率が低い可能性がある．一方，性別予測をマーケティングやオピニオンマイニングに応用することを考えると，著名人以外の一般ユーザの性別を正確に推定することの意義は大きい．そのため，一般ユーザに対する

性別のラベル付けによるデータ拡張を行い，性別推定の分類器の一般ユーザに対する性能を向上させる．

データ拡張の処理の流れを図 3.15 に示す．まず最初に，ラベル付きデータによって性別推定の分類器を学習する．次に，主に一般の Twitter ユーザの情報を大量に取得し，ラベルなしデータとして用意する．学習した分類器を用いてラベルなしデータにおけるユーザの性別を予測する．性別判定の信頼度が高いとき，すなわち性別スコアが 1(男性) または 0(女性) に近い値の時，そのユーザに予測した性別ラベルを付与し，ラベル付きデータに追加する．ラベルなしデータに対して性別を予測する方式として,「自己学習によるデータ拡張」と「統合モデルによるデータ拡張」の 2 つを用いる．

以下，3.3.1 項ではラベルなしデータの収集について述べる．その後 3.3.2 項では自己学習によるデータ拡張，3.3.3 項では統合モデルによるデータ拡張について説明する．

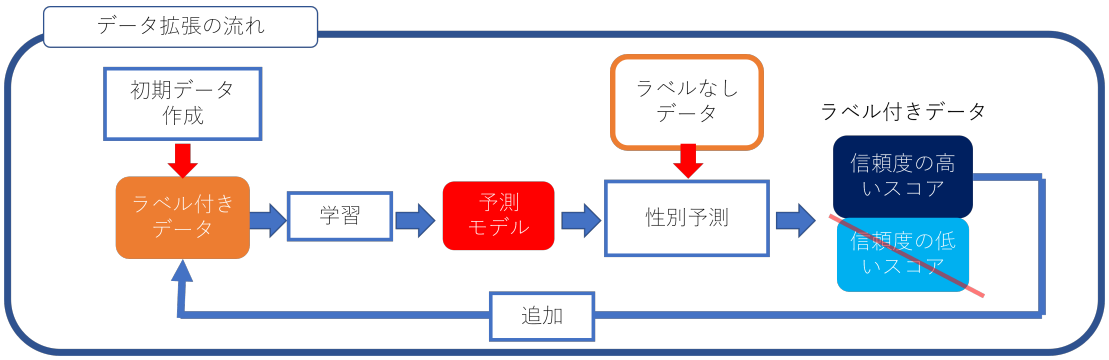

図 3.15: データ拡張の概要

3.3.1 ラベルなしデータ

ラベルなしデータを用意するため，Twitter ユーザを新たに選別し，その情報を収集する．まず，3.1 節で述べたデータセットの構築と同じように，Twitter 日本フォロワー数総合ランキング [14] のウェブサイトからラベルなしデータとして収集するユーザを選別する．3.1 節でのデータセット構築時には上位 2 万件のユーザを取得したが，それとの重複を避けるため，フォロワー数のランキングの上位 2 万から 3 万のユーザを選ぶ．フォロワー数が比較的少ないため，この中には著名人以外の一般のユーザが多く含まれると考えられる．

次に，Twitter API によって分類器の学習に必要な情報の取得を試みる．ツイートについては，3.1 節のデータセットと同じく，そのユーザの最新の 100 件のツイート (テキストと画像) の取得を試みる．これに成功したユーザをラベルなしデータに加える．最終的に 6,845 名のユーザの情報が得られた．

3.3.2 自己学習によるデータ拡張

4 章で述べる評価実験では，本章で紹介した複数の分類器のそれぞれについて，データ拡張の効果を検証する．ラベルなしデータに対して性別ラベルを付与する際，評価対象の分類器を用いて性別スコアを予測する．すなわち，ある分類器の訓練データを拡張する際には，初期のラベル付きデータを用いて学習したその分類器を用いて，ラベルなしデータに対する性別推定を行う．

性別推定の信頼度を式 (3.1) のように定義する．

$$Reli(u) = 1 - min(|1 - s(u)|, |0 - s(u)|) \tag{3.1}$$

ここで u はユーザ，$s(u)$ はそのユーザに対する性別スコア (0 から 1 までの値) である．すなわち，性別スコアと 1 または 0 の差が小さいとき，性別推定の信頼度を高く見積る指標となっている．全てのラベルなしデータに対してユーザの性別スコアを求めた後，判定の信頼度が大きい上位 3,000 件のユーザを選別し，これを初期の訓練データに加える．その後，初期のデータと自動拡張したデータを合わせて訓練データとし，分類器を再度学習する．

3.3.3 統合モデルによるデータ拡張

もうひとつのデータ拡張の手法として，ラベルなしデータに対する性別を推定する際に，統合モデルを用いる．統合モデル以外の分類器について，データ拡張の効果を測るときも，性別スコアの計算に統合モデルを用いる．先ほどと同様に，判定の信頼度が高い上位 3000 件のユーザを初期の訓練データに追加し，分類器を再学習する．この方式では，個々の分類器についてデータ拡張の効果を測る際，使用される自動拡張データは全て同じである．

ここでの目的は，拡張データとして比較的質の高いデータを構築し，これを用いることで分類器の性別推定の正解率がどれだけ向上するかを検証することである．理想的なデータ拡張とは，正しい性別ラベルが付与されたデータを追加することである．一方，前述の自己学習によるデータ拡張では，初期のラベル付きデータによって学習された分類器の精度が低い場合，拡張データの質が悪くなることが予想される．訓練データを増やすことによって，分類器の性能がどれだけ向上するかを主に検証するためには，理想的なデータ拡張に近い条件で自動ラベル付けデータを構築する方が望ましい．4.3 節で報告するように，統合モデルの正解率は他のどの分類器よりも高かった．そのため，統合モデルによって自動ラベル付けを行うことで，比較的質の高いラベル付きデータを獲得する．

第4章　評価実験

本節では提案手法の評価実験について述べる．4.1 節では，実験のために構築したデータセットについて述べる．4.2 節では，ツイートを対象に性別を判定する分類器を評価する．4.3 節では，Twitter の様々な情報を素性とする分類器を評価する．また，データの自動拡張によって性別推定の性能がどの程度向上するかを検証する．

4.1　実験データ

本節では，評価実験のために作成したデータセットについて，その構築方法やデータ数などを報告する．実験では 3 種類のデータセットを用いる．以下，順に説明する．

4.1.1　初期データ

「初期データ」とは，最初の分類器の学習に使う性別ラベル付きの Twitter ユーザの集合を指す．3.1 節に述べた方法で作成した．すなわち，著名人の Twitter アカウントを中心にユーザを収集し，その性別ラベルを自動的に付与した．

初期データの統計情報を表 4.1 の「初期データ」の列に示す．取得したユーザ数，ツイートの総数，プロフィール画像の総数，ヘッダー画像の総数，ならびにこれらの男性と女性の内訳を掲載した．結果として，男性 2,107 名，女性 2,893 名，計 5,000 名のデータを獲得した．また，1 ユーザに対して最大 100 件のツイートを収集した結果，合計で 383,201 件のツイートを取得した．

後述する一部の実験では，これを 8:1:1 の割合で訓練，開発，テストデータに分割して使用した．

4.1.2　ラベルなしデータ

「ラベルなしデータ」とは，3.3 節で述べたデータ拡張の際に用いるデータである．すなわち，性別ラベルの付いていない一般ユーザの集合であり，これに自動的にラベルを付与することで訓練データを拡張する．このデータセットは，3.3.1 項で説明した方法で構築した．

ラベルなしデータの統計情報を表 4.1 の「ラベルなしデータ」の列に示す．ユーザ数は 6,845 名，ツイート数は合計で 560,274 件である．なお，性別ラベルは付与されていないので，男女の内訳は不明である．

表 4.1: データセットの統計

	初期データ	ラベルなしデータ	評価データ
ユーザ数	5,000	6,845	459
男性	2,107	–	286
女性	2,893	–	173
ツイート数	383,201	560,274	33,083
男性	156,316	–	20,654
女性	226,885	–	12,429
プロフィール画像数	3,704	6,075	442
男性	1,653	–	276
女性	2,051	–	166
ヘッダー画像数	3,438	4,617	378
男性	1,396	–	229
女性	2,042	–	149

4.1.3 評価データ

前述の初期データは，自動的に性別ラベルが付与されており，誤りを含む可能性がある．性別推定モデルの性能を厳密に評価するためには，人手でラベル付けされたデータセットが必要である．そのため比較的少数のユーザに対して，人手で性別のラベルを付与し，評価データを構築する．

既に述べたように，本研究では著名人以外の一般ユーザの性別を推定することを重視しているため，評価データに用いるユーザも著名人ではない一般の人のユーザを選ぶ．初期データの著名人ユーザをフォローしているユーザのうち，フォロワー数が 1,000 人以下のユーザを選別する．フォロワー数に制限を設けたのは，フォロワー数が多いユーザは著名人の可能性が高く，これを除外するためである．最終的に 500 名のユーザを選定した．

次に，3 名の被験者に，Twitter ユーザの最新の 20 件のツイートやプロフィール画面を参照し,「男性」「女性」「不明」のいずれかを付与することを依頼する．またこの作業を実施するためのウェブ上のインターフェースを開発した．このインターフェースによる作業画面を図 4.1 に示す．画面の右半面は対象ユーザのタイムライン (ユーザが投稿したツイート一覧) を埋め込んで表示させている．画面の左

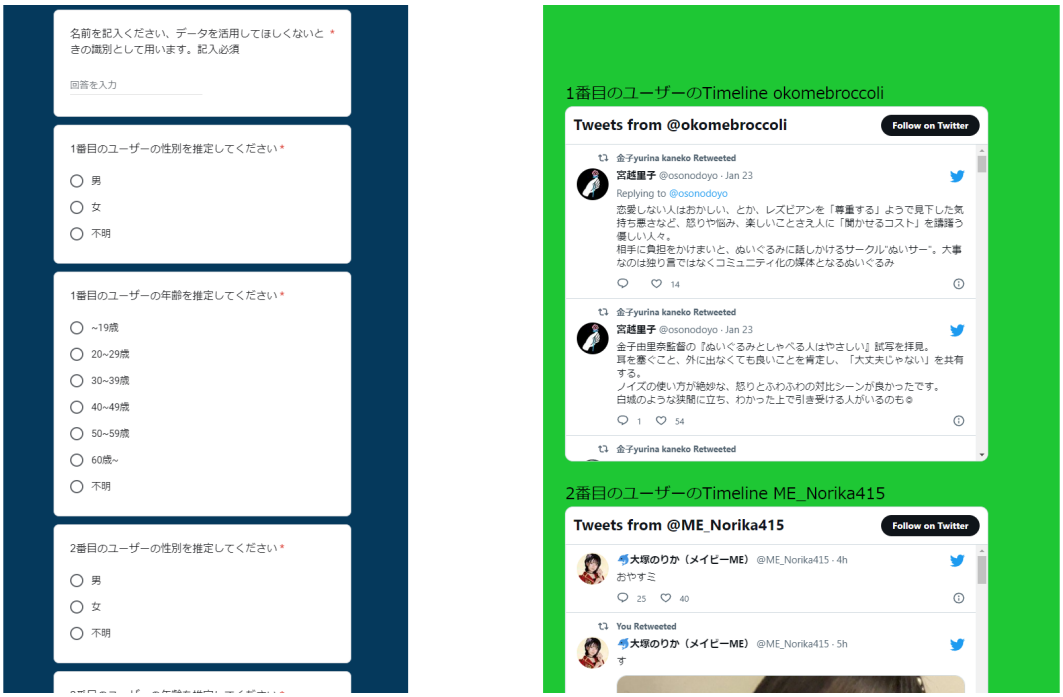

図 4.1: ラベル付けのためのインターフェース

半面は被験者が右のユーザの性別や年齢を推定し，該当する結果を入力する画面である．なお，今回のラベル付け作業では，ユーザのプロフィールとして性別と年齢を推定して記入することを依頼したが，今回の実験では年齢の情報は使用しない．

複数人の被験者が付与したラベルがどれだけ一致しているかを確認するため，作業結果の Fleiss の κ 係数を算出した．Fleiss の κ 係数の定義を式 (4.1) から (4.4) に示す．Fleiss の κ は，作業者が 2 名ではなく 3 名以上のときに，作業結果の一致度を測るために用いられる．$\bar{P}$ は，付与したラベルの実際の一致度，$\bar{P}_e$ はその期待値を表す．n_{ij} は，i 番目の評価対象 (データ) に j 番目のラベルを付与した被験者の数，N, n, k はそれぞれ評価対象の数，被験者の数，ラベルの数を表す．計算の結果，κ 係数は 0.747 となり，比較的高い一致率で性別ラベルが付与されたことを確認した．

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{4.1}$$

$$\bar{P} = \frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^2 - Nn\right) \tag{4.2}$$

$$\bar{P}_e = \frac{1}{N}\left(\sum_{j=1}^{k} p_j^2\right) \tag{4.3}$$

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \tag{4.4}$$

2 名以上の被験者が「男性」または「女性」のラベルを付与したとき，そのユーザを評価データに加えた．最終的に，男性 286 名，女性 173 名，計 459 名からなる評価データを構築した．評価データの詳しい統計情報は，表 4.1 の「評価データ」の列に示す．ツイートは 1 ユーザにつき最大で 100 件取得しているが，実際に取得できたツイート数は 1 人当たりの平均で 72.1 ツイートである．プロフィール画像を取得できたユーザの割合は 96.3%で，ほとんどのユーザがプロフィール画像を設定している．一方，ヘッダー画像を取得できたユーザの割合は 82.4%となり，プロフィール画像より低い．

4.2 ツイートによる分類器の評価

本節では，3.2.2 項で述べたツイートによる分類器を評価する．この分類器は，(1) 個々のツイートの性別スコアの推定，(2) ユーザの性別スコアの推定，という 2 段階でユーザの性別を推定する．個々のツイートの性別スコアを推定する分類器として，Text only model，Image only model，Early fusion model，Late fusion model，Dense fusion model，Caption model，Ensemble model を用い，これらを比較する．次に，3.2.2 項の冒頭で説明したように，個々のツイートの性別スコアの平均を取る Tw-ave，Softmax 関数を用いて正規化した性別スコアの平均を取る Tw-soft，性別スコアが 1 または 0 に近い上位 10%のツイートの性別スコアの平均を取る Tw-sel の 3 つの方式により，ユーザの性別スコアを計算する．このスコアが 0.5 以上であれば男性，0.5 未満であれば女性と判定し，これを正解の性別クラスと比較し，性別推定の正解率を算出する．分類器の学習に用いるデータは，4.1.1 項で述べた初期データを分割して作成された訓練データと開発データを使用する．分類器を評価するために用いるデータは，初期データを分割して作成されたテストデータ，ならびに 4.1.3 項で述べた評価データを用いる．実験結果を表 4.2 に示す．

本研究は，テキストと画像を同時に考慮することを特徴とした分類器を複数提案している．これらの分類器は画像付きのツイートに対して有効に働くことを期待している．一方，ツイートの中には，画像を含まずテキストだけで投稿されたものも多い．テキストのみのツイートに対して，テキストと画像の同時利用を想定したモデルは，性別を正確に予測できないことも考えられる．そこで，画像付きのツイートのみを利用してユーザの性別を推定する実験を行う．

データセットは，初期データを訓練，開発，テストデータに分割したものを用いる．分類器の学習時には，訓練データおよび開発データから，画像を含まないツイートを削除し，画像付きのツイートだけのデータを作成し，これを用いて分類器を学習する．テストの際も同様に，テストデータから画像を含まないツイートを削除する．また，画像付きツイートを 1 件も投稿していないユーザについては，

表 4.2: ツイートによる分類器を用いた性別推定の正解率

Model	初期データ			評価データ		
	Tw-ave	Tw-soft	Tw-sel	Tw-ave	Tw-soft	Tw-sel
Text only	.854	.852	.845	.659	.667	.624
Image only	.776	.774	.762	.754	.758	.754
Early fusion	.791	.791	.786	.732	.739	.721
Late fusion	.654	.608	.561	.617	.617	.725
Dense fusion	.582	.582	.672	.730	.732	.708
Caption	.847	.855	.823	.810	**.813**	.754
Ensemble	.878	.881	**.893**	.697	.697	.699

これをテストデータから除外する．これに該当するユーザは 38 名で，全体 (500 名) の 7.6%を占める．すなわち，この実験のテストデータのユーザ数は 462 名である．個々のツイートの性別スコアを予測する分類器として，Text only model，Image only model，Early fusion model，Late fusion model，Caption model を用いた．Dense fusion モデルは表 4.2 に示すように正解率が低く，学習に時間がかかるため，比較の対象とはしなかった．個々のツイートの性別スコアからユーザの性別スコアを計算する方式として，Tw-ave と Tw-soft を用いる[1]．実験の結果を表 4.3 に示す．

表 4.3: 画像付きツイートのみを利用した分類器による性別推定の正解率

Model	初期データ	
	Tw-ave	Tw-soft
Text only	.802	.796
Image only	.883	.882
Early fusion	.893	.893
Late fusion	.688	.688
Caption model	.893	.892

実験結果を考察する．表 4.2 に示したように，初期データでは，Text only モデルの正解率が，テキストと画像の両方の情報を用いる Early fusion，Late fusion，Dense fusion の各モデルよりも高い．ただし，表 4.3 に示したように，画像を含むツイートのみを利用したときは，Early fusion モデルの精度が一番高いことが確認された．これらの事実が予備実験により確認できたため，Text only model と Early fusion model を組み合わせた Ensemble model を考案している．実験の結果，Ensemble モデルの最高の正解率は 0.893 となり，他のどの分類器よりも正解率が高いことが確認された．なお，Tw-ave，Tw-soft，Tw-sel の優劣はモデルによって

[1]Tw-sel はシステムの不具合のため比較の対象としなかった．

異なり，どの手法が最適であるかを結論付けることはできない．

一方，評価データでは，Image only model は Text only model よりも正解率が高く，Caption model 以外のテキストと画像を同時に使用するモデルよりも高い．著名人ではない一般ユーザの場合，男女によって投稿する画像に顕著な違いがあり，テキストよりも画像の方が性別判定の有力な手がかりになっている可能性がある．例外は Caption モデルで，比較した分類器の中では最高の成績を収めている．また，初期データの実験結果を見ると，テキストと画像を同時に考慮した単独の (Ensemble モデル以外の) 分類器の中では，Caption モデルの正解率が最も高い．画像を埋め込みベクトルに変換するのではなく，画像の内容を表すテキストに変換してから，ツイート本文とともに性別を予測すると正解率が向上するのは興味深い．

初期データと評価データの実験結果の違いを考察する．初期データを用いた実験では，Ensemble モデルを除く単独の分類器では Text only モデルの正解率が最も高いことから，画像よりもテキストの情報が性別判定に有力であると言える．一方，評価データでは，Text only モデルの正解率は低く，テキストよりも画像の情報が性別判定に有力であると言える．初期データは主に著名人のユーザであり，性別ラベルは自動的に付与されているのに対し，評価データは主に一般のユーザであり，性別ラベルは人手で付与されている．初期データと評価データとでテキスト素性と画像素性の有効性が異なるのは，上記のようなデータセットの違いに起因すると考えられるが，はっきりとした因果関係はわかっていない．これを精査することは今後の課題である．

4.3 Twitter の情報による分類器の評価

4.3.1 性別判定分類器の評価

本節では 3.2.2 項から 3.2.8 項で述べた分類器の性能を評価する．分類器の学習に用いるデータとして，4.1.1 項で述べた初期データを分割して得られた訓練データと開発データを使用する．分類器の評価に用いるデータとして，4.1.3 項で述べた評価データを用いる．先程の実験と同様に，各分類器が推測する性別スコアが 0.5 以上であれば男性，0.5 未満であれば女性と予測した際の正解率を評価基準とする．結果を表 4.4 に示す．Tw-というプリフィックスで示された分類器は，ツイートによる分類器のうち，表 4.2 に示す初期データを用いた予備実験で最も正解率の高かった Ensemble model を用いている．

統合モデルの正解率は 0.851 であり，個々の分類器よりも高い．ツイートのテキストや画像だけでなく，Twitter から得られる様々な情報を考慮することが性別推定の正解率の向上に寄与することが確認できる．個々の分類器では，プロフィール画像を用いた分類器の正解率が最も高い．評価データ内のおよそ 100 名のユーザについて，そのプロフィール画像を確認したところ，69%のユーザが一人の人物

のみが映った画像であることが確認された．そのため，プロフィール画像はユーザ本人の画像である可能性が高い．また，被写体が人間でなくても，デフォルメされたユーザ本人と見受けられるキャラクターであったり，好きなキャラクター，動物の画像などが使われており，このような画像もユーザの特徴を十分に表しているといえる．以上の実験結果並びに考察から，プロフィール画像は性別推定の有力な情報であることが確認された．それに対し，ヘッダー画像としては風景，ロゴ，集合写真，ゲームの画面などが使われることが多く，プロフィール画像と比べてユーザ本人の情報を直接開示している可能性が低い．そのためヘッダー画像のみを用いた分類器の正解率は 0.599 と低かった．また，ユーザ名を用いた分類器も正解率が高く，名前に性差が強く表れていることを示唆する．フォロイーの自己紹介文による分類器は，比較した分類器の中では正解率が最も低かった．単語を素性とするいわゆる bag-of-words モデルで素性ベクトルを作成したため，テキストの文脈は分類器に反映されない．そのため，文脈を考慮したモデルを用いることで，フォロイーの紹介文を用いた分類器の性能が向上する可能性がある．例えば，BERT を用いてそれぞれのフォロイーの自己紹介文を特徴ベクトルに変換し，その平均ベクトルを素性とした分類器を学習する事が考えられる．

統合モデルは Light GBM によって分類器を学習している．Light GBM では，学習に使用した個々の素性の重要度を計算することができる．大まかに言えば，Light GBM は弱分類器として決定木が使われているが，決定木におけるデータ分岐の際に使われた素性に関する質問の回数によって素性の重要度を定義している．統合モデルにおける個々の分類器の重要度を表 4.5 に示す．重要度が高いのは，Twitter 統計量による分類器，フォロイーの自己紹介文による分類器，自己紹介文による分類器である．これらの分類器は性別予測の正解率が特に高いわけではないが，統合モデルにおける重要度は高い．この理由として，他の分類器では考慮されていない独自の特徴が捉えられており，性別推定の正解率の向上に大きく寄与したと考えられる．

表 4.4: 個々の分類器ならびに統合モデルの評価データに対する性別推定の正解率

Tw-ave	Tw-soft	Tw-sel	統計	ユーザ名	画像-P	画像-H	画像-PH	自己	フォ自己	統合
.697	.697	.699	.636	.764	.810	.599	.643	.656	.584	.851

(分類器の略号) 統計:Twitter の統計量，画像-P: プロフィール画像，
画像-H: ヘッダー画像，画像-PH: プロフィール&ヘッダー画像，
自己: 自己紹介文，フォ自己: フォロイーの自己紹介文

表 4.5: 統合モデルにおける個々の分類器の重要度

Tw-ave	Tw-soft	Tw-sel	統計	ユーザ名	画像-P	画像-H	画像-PH	自己	フォ自己
296	295	232	363	465	284	211	214	342	363

4.3.2 データ拡張の評価

本項では，3.3 節で述べたデータ拡張の効果を評価する．前節で評価した 11 個の分類器について，以下に示す異なる訓練データを用いて分類器を学習する．

初期データ 初期データのみを訓練データとする場合．

初期+拡張 S 初期データに加え，3.3.2 項で述べた自己学習によるデータ拡張によって得られた自動ラベル付きデータも訓練データとして用いる場合．「S」は自己学習 (Self-training) を表す．

初期+拡張 I 初期データに加え，3.3.3 項で述べた統合モデルによるデータ拡張によって得られた自動ラベル付きデータも訓練データとして用いる場合．「I」は統合モデル (Integrated Model) を表す．

訓練データで分類器を学習した後，それを用いて評価データのユーザの性別を推定し，正解率を求める．データ拡張によって正解率が向上するか，自己学習によるデータ拡張と統合モデルによるデータ拡張のどちらが効果的かを検証する．

実験結果を表 4.6 に示す．訓練データが「初期データ」の行は，表 4.4 の実験結果の再掲である．「初期+拡張 S」と「初期+拡張 I」は上記の実験設定に対応する．なお，統合モデルは，「初期+拡張 S」と「初期+拡張 I」の実験条件とで同じモデルが得られる．

表 4.6: データ拡張を用いて学習された分類器の性別推定の正解率

訓練データ	Tw-ave	Tw-soft	Tw-sel	統計	ユーザ名	画像-P	画像-H	画像-PH	自己	フォ自己	統合
初期データ	.697	.697	.699	.636	.764	.810	.599	.643	.656	.584	.851
初期+拡張 S	.656	.719	.599	.644	.769	.755	.553	.603	.664	.612	.806
初期+拡張 I	.749	.763	.736	.694	.795	.808	.577	.675	.699	.597	.806

(分類器の略号) 統計:Twitter の統計量，画像-P: プロフィール画像，
画像-H: ヘッダー画像，画像-PH: プロフィール&ヘッダー画像，
自己: 自己紹介文，フォ自己: フォロイーの自己紹介文

「初期+拡張 S」の結果を見ると，拡張データを使用することで，個々の分類器では正解率が向上したものもあるが，統合モデルの正解率は向上しなかった．特に正解率の高かったプロフィール画像を用いた分類器の性能が大きく低下していることから，統合モデルの正解率が 0.045 ポイント下がった．

「初期+拡張 I」の結果を見ると，初期データのみを訓練データとした場合と比べて，多くの分類器について正解率が向上したことが確認できた．また,「初期+拡張 S」では正解率が大きく低下したプロフィール画像による分類器の正解率は,「初期+拡張 I」ではそれほど大きくは低下していない．しかし，統合モデルの正解率はデータ拡張によって改善しなかった.「初期+拡張 S」と「初期+拡張 I」を比較すると，ほとんどの分類器で後者の方が正解率が高い．後者では，拡張データを得る手法として，それぞれのモデルの代わりに正解率の高い統合モデルの判定結果に基づいてラベルを決めているため，この結果は自然である．このことから，今回の実験では統合モデルの正解率はデータ拡張によって向上しなかったが,自動拡張データの品質が向上すれば，訓練データの増加が正解率の向上に寄与する可能性がある．また，今回の実験では，個々の分類器で追加する事例数を同じにするために，判定の信頼度の上位 3000 件のデータを追加したが，ある閾値以上の信頼度の事例のみ追加するなど，自動拡張するデータの性別ラベルの誤りを減らす工夫が必要である．

データ拡張が正解率の向上につながらなかった原因を探るため，実験結果のより詳細な分析を行う．表 4.7 は，自己学習によるデータ拡張によって，訓練データとして追加されたユーザの性別スコアの 0 または 1 に近い閾値を示している．本研究のデータ拡張では，式 (3.1) に示した判定の信頼度が高い上位 3000 件のユーザに性別ラベルを付与して訓練データに追加している．この 3000 件のユーザの性別スコアのうち，最も 0 に近いもの，1 に近いものを調べ，これをデータを追加するか否かと閾値としている．ツイートによる分類器 (Tw-sel) やヘッダー画像による分類器は，表 4.7 の閾値が 0.5 に近いが，これらの分類器はデータ拡張によって正解率が向上しなかった．これは，性別スコアが 0.5 付近である，判定が曖昧なユーザが拡張データに追加され，このことが分類器の正解率の低下の原因になっていると考えられる．しかし，フォロイーの自己紹介文による分類器も閾値が 0.5 に近いが，データ拡張により正解率は向上している．また，プロフィール画像による分類器は，閾値が比較的 0 または 1 に近いのにも関わらず，正解率は低下している．これらの原因ははっきりしない．

表 4.7: 拡張データに追加されたユーザの性別スコアの閾値

	Tw-ave	Tw-soft	Tw-sel	統計	ユーザ名	画像-P	画像-H	画像-PH	自己	フォ自己	統合
0 付近	.341	.366	.445	.364	.277	.207	.426	.393	.018	.311	.053
1 付近	.659	.633	.555	.635	.722	.793	.574	.606	.981	.689	.946

(分類器の略号) 統計:Twitter の統計量，画像-P: プロフィール画像，
画像-H: ヘッダー画像，画像-PH: プロフィール&ヘッダー画像，
自己: 自己紹介文，フォ自己: フォロイーの自己紹介文

図 4.2 は，それぞれの分類器で評価データのユーザの性別を予測したときの性別スコアの分布を示している．これを見ると，分類器によって，性別スコアは 0 または 1 付近に極端に分かれるものとそうでないものがあることがわかる．例えば,

ツイートによる分類器 (Tw-ave) では，データ拡張後において，女性と判定するときの性別スコアは 0 に近いものが多いのに対し，男性と判定するときの性別スコアはそれほど 1 付近に偏っていない．このことから，ツイートには男性らしさよりも女性らしさを示唆する情報が多いことが推察できる．また，自己紹介文による分類器によって予測される性別スコアは 0 または 1 付近に極端に偏っているが，正解率はそれほど高くないことから，男性 (または女性) を高い信頼度で女性 (または男性) に誤判定していると言える．初期データと評価データとで自己紹介文の内容に大きな差がある可能性もある．

全体的に，データ拡張を行うことで，性別スコアの分布が 0 または 1 に偏る傾向が見られる．このことは，データ拡張によって正解率が改善した分類器については，その裏づけとなる．また，プロフィール画像による分類器は，データ拡張によって正解率が向上しなかったが，図 4.2 を確認すると，初期データと比べて性別スコアの分布に極端な偏りが見られない．したがって，性別スコアの分布の偏りと分類器の正解率にはある程度の相関があることが認められる．

データ拡張によって正解率が改善しない原因として，追加されるユーザの男女比の偏りが考えられる．追加されるユーザの多くが男性または女性の場合，データセット全体で性別クラスのバランスが崩れ，分類器の正解率の低下を招く原因となりうる．このことを検証するため，自己学習によるデータ拡張の際に追加された男性ユーザと女性ユーザの数を調査した．結果を表 4.8 に示す．その結果，男性が多く追加される分類器もあれば，女性が多く追加される分類器もある．しかし，追加される男性ユーザと女性ユーザの数に大きな差は見られなかった．

表 4.8: 自己学習によるデータ拡張によって追加された男女数

性別	Tw-ave	Tw-soft	Tw-sel	統計	ユーザ名	画像-P	画像-H	画像-PH	自己	フォ自己	統合
女性	1905	1905	1888	1899	1505	1057	1546	1074	1788	1206	1888
男性	1095	1095	1112	1101	1495	1943	1454	1926	1212	1794	1112

(分類器の略号) 統計:Twitter の統計量，画像-P: プロフィール画像，
画像-H: ヘッダー画像，画像-PH: プロフィール&ヘッダー画像，
自己: 自己紹介文，フォ自己: フォロイーの自己紹介文

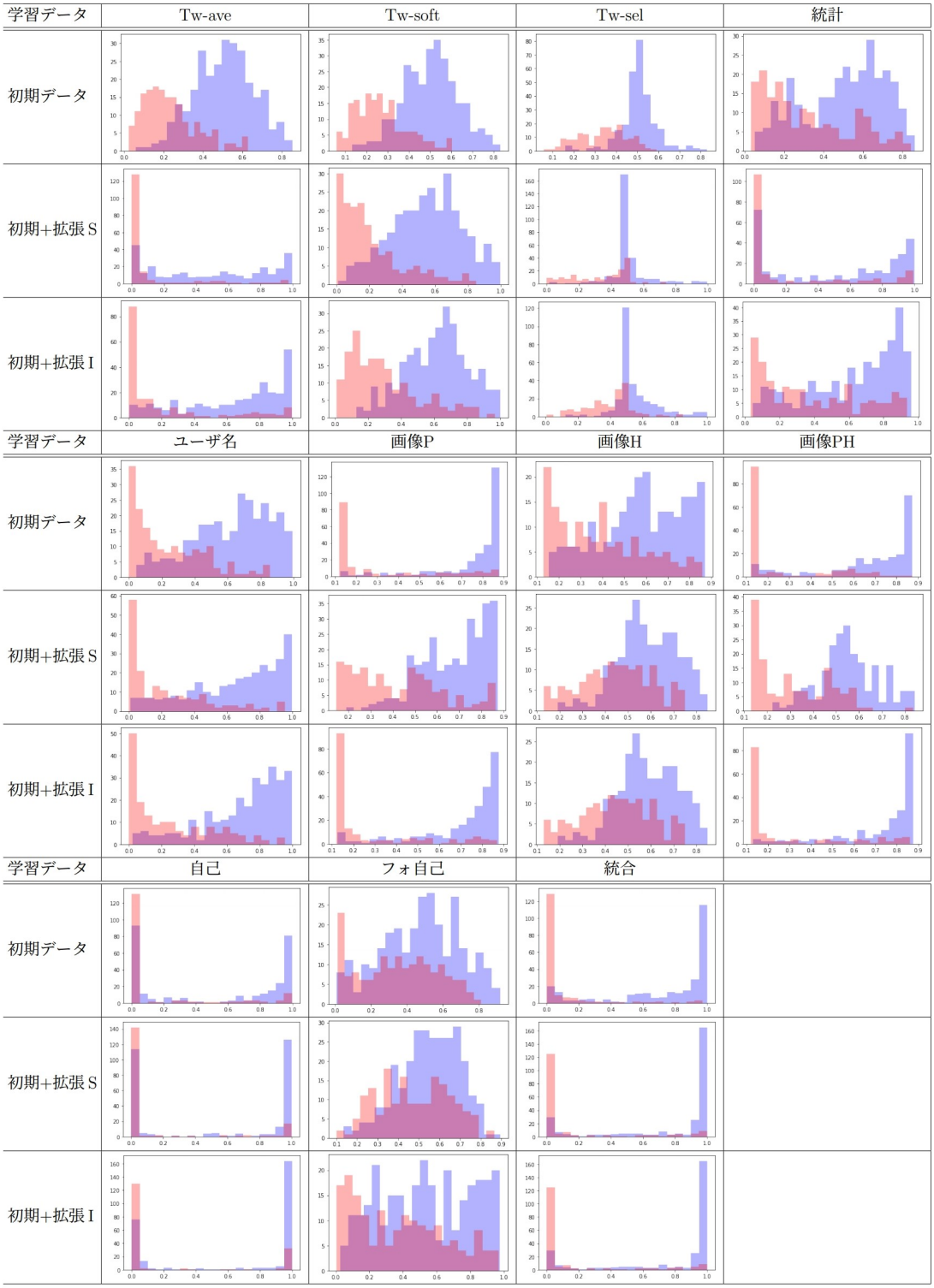

図 4.2: 分類器によって予測された性別スコアの分布

第5章　おわりに

5.1　本研究のまとめ

本論文では，主に一般の Twitter ユーザの性別を予測することを目的とし，多様な特徴量を考慮した分類器を学習する手法を提案した．性別推定の分類器を学習するためには，性別のラベルが付与されたデータセットが必要であるため，まず，主に著名人の Twitter ユーザに対し，Wikipedia などの外部情報を利用して性別の自動ラベル付けを行い，性別ラベル付きのユーザのデータセットを構築した．次に，これを訓練データとして性別を推定する分類器を学習した．機械学習のための素性として，ツイート，Twitter 統計量，ユーザ名，プロフィール画像，ヘッダー画像，自己紹介文，フォロイーの自己紹介文を用いた．特にツイートを素性とする分類器については，ユーザが投稿したツイートのテキストと画像を同時に考慮する 5 つの分類器を提案した．また，これらの分類器が出力する性別の予測スコアを素性とし，複数の分類器の情報を参照して性別を推定する統合モデルを提案した．また，上記の手続きで構築したラベル付きデータ (初期データと呼んだ) は訓練データとして十分な量がない可能性があること，著名人のユーザにラベル付けしたために一般のユーザに対する性別推定の正解率が低下する懸念があることから，初期データから学習した性別推定の分類器を用いて一般ユーザの性別ラベルを自動的に付与することにより，訓練データを自動拡張する手法を提案した．

提案手法の評価のため，初期データを訓練，開発，テストデータに分割して分類器の正解率を測る実験と，500 名程度のユーザに人手で性別のラベルを付与して作成した評価データを用いて分類器の正解率を測る実験を行った．まず，ツイートを素性とした 7 つの分類器 (2 つのベースラインを含む) の性別推定の正解率を測り，それらを比較した．初期データを用いた実験では，Text only model と Early fusion model を組み合わせた Ensemble model が最も正解率が高く，その値は 89.3%であった．一方，評価データを用いた実験では，Caption model の正解率が 78.9%と最も高かった．初期データと評価データで最高の正解率が得られた分類器は異なるが，ツイートと画像を同時に考慮して分類器を学習するアプローチの有効性が確認された．

次に，Twitter の様々な情報を用いた分類器ならびに統合モデルを評価した．評価データに対する性別推定で最も高い正解率が得られたのは統合モデルであり，その値は 85.1%であった．多様な特徴量を考慮することで性別推定の性能が向上することがわかった．

データ拡張の評価実験では，正解率が最も高い統合モデルについては，データ拡張で訓練データを増やしてもその正解率は向上しなかった．ただし，個々の分類器の中にはデータ拡張によって正解率が向上するものもあり，データ拡張が性別推定の性能向上に貢献する可能性を示した．

5.2 今後の課題

評価実験では提案手法の有効性が確認できたものの，データ拡張による正解率の向上が確認できたのは限られた分類器のみであること，個々の分類器でも正解率が十分に高くないものがあることなど，改善が必要な点もいくつか見つかった．これを踏まえ，今後の課題を以下にまとめる．

データ拡張手法の改善　今回の実験では，データ拡張の際，判定の信頼度が高い上位 3,000 件のデータを追加した．これは，複数の分類器についてデータ拡張の効果を公平に比較するために，同数のデータを訓練データに追加するようにしたためである．しかし，この方法では判定の信頼度が低いデータを訓練データとして追加し，結果として訓練データに誤りが多く含まれるようになった可能性がある．そのため，ある閾値以上の信頼度の事例のみをラベル付けし訓練データに追加することで，ノイズとなるデータを減らす方法が考えられる．

また，データ拡張によってプロフィール画像による分類器の正解率が大きく低下したが，この原因は不明なままである．今後，この原因を精査し，正解率低下の要因を取り除くことができれば，プロフィール画像による分類器だけでなく統合モデルの正解率の改善も見込まれる．

分類器の改良　フォロイーの自己紹介文を素性とする分類器を作成した際，自己紹介文に含まれる単語を素性，その TF-IDF 値を値とする特徴ベクトルを素性として分類器を学習した．このため，単語の並びなどの文脈に関する情報は考慮されていなかった．そこで，文脈を考慮した文の埋め込み表現を出力できる BERT を用いて，フォロイーの自己紹介文を埋め込み表現 (特徴ベクトル) に変換し，複数のフォロイーに対してその平均ベクトルを求めてユーザの特徴ベクトルとすることで，文脈を考慮した分類器を学習する．

また，ツイートによる分類器や自己紹介文による分類器を学習する際，ツイートや自己紹介文に含まれる絵文字を除外した．しかし，男性と女性とで使用する絵文字に違いがあると考えられることから，絵文字も素性として利用することを検討する必要がある．

実験では，Twitter 統計量による分類器，フォロイーの自己紹介文による分類器，自己紹介文による分類器については，正解率が高いわけではなかったが，統合モデルにおける重要度は高かった．このことから，これらの分類器は他の分類器では捉えられていない独自の特徴が反映されていると考えられる．今後，そのよう

な特徴が何かを分析し，これを明示的に素性として利用したり効率的に学習したりすることができる手法を探究する．

性別以外のプロフィールの予測　本研究では，ユーザプロフィールのうち性別に注目し，これを自動的に推定する手法を探究した．今後は，年齢や出身地といった他のプロフィールの自動推定に提案手法を適用し，その有効性を検証したい．

参考文献

[1] Googletrans, (2023-1 閲覧). https://pypi.org/project/googletrans/.

[2] Pillow, (2023-1 閲覧). https://pillow.readthedocs.io/en/stable/.

[3] Radford Alec, Wook Kim Jong, Hallacy Chris, Ramesh Aditya, Goh Gabriel, Agarwal Sandhini, Sastry Girish, Askell Amanda, Mishkin Pamela, Clark Jack, Krueger Gretchen, and Sutskever Ilya. Learning transferable visual models from natural language supervision, (2021).

[4] Dosovitskiy Alexey, Beyer Lucas, Kolesnikov Alexander, Weissenborn Dirk, Zhai Xiaohua, Unterthiner Thomas, Dehghani Mostafa, Minderer Matthias, Heigold Georg, Gelly Sylvain, Uszkoreit Jakob, and Houlsby Neil. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[6] Face++, (2022-2 閲覧). https://www.faceplusplus.com/.

[7] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, pp. 1189–1232, 2001.

[8] Yaguang Liu, Lisa Singh, and Zeina Mneimneh. A comparative analysis of classic and deep learning models for inferring gender and age of Twitter users. In *Proceedings of the 2nd International Conference on Deep Learning Theory and Applications*, pp. 48–58, 2021.

[9] Xiaojun Ma, Yukihiro Tsuboshita, and Noriji Kato. Gender estimation for SNS user profiling using automatic image annotation. In *IEEE International Conference on Multimedia and Expo Workshops*, pp. 1–6, 2014.

[10] Antonio A. Morgan-Lopez, Annice E. Kim, Robert F. Chew, and Paul Ruddle. Predicting age groups of Twitter users based on language and metadata features. *PLOS ONE*, Vol. 12, No. 8, p. e0183537, 2017.

[11] Shigeyuki Sakaki, Yasuhide Miura, Xiaojun Ma, Keigo Hattori, and Tomoko Ohkuma. Twitter user gender inference using combined analysis of text and image processing. In *Proceedings of the 25th International Conference on Computational Linguistics*, pp. 54–61, 2014.

[12] Jae-Wook Seol, Kwang-Yong Jeong, and Kyung-Soon Lee. Follower classification through social network analysis in Twitter. In *Grid and Pervasive Computing*, pp. 926–931, 2013.

[13] 杉谷卓哉, 白川真澄, 原隆浩, 西尾章治郎. 教師あり機械学習を用いたツイート投稿時のユーザ位置推定手法. 情報処理学会研究報告, Vol. 2013-DBS-158, No. 26, pp. 1–8, 2013.

[14] Twitter 日本 フォロワーランキング, (2022-12 閲覧). `https://meyou.jp/ranking/follower_allcat`.

[15] Zijian Wang, Scott A. Hale, David Adelani, Przemyslaw A. Grabowics, Timo Hartmann, Fabian Flöck, and Dacid Jurgens. Demographic inference and representative population estimates from multilingual social media data. In *Proceedings of the World Wide Web Conference*, pp. 2056–2067, 2019.