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.