A Scalable Holistic approach for Age and Gender inference of Twitter Users

Anonymous ACL submission

Abstract

Numerous studies have focused on inference of age and gender. We consider a new approach that takes advantage of contrastive learning methods by using both text and image content for this prediction task. We also consider the case where only text or image data is available. Under both of these conditions, we show that our model achieves better performance than the state-of-the-art ones, and still performs well with text/images only. Moreover, because demographic datasets can be small, we also consider combining different datasets to understand when augmentation is valuable and when it is not.

1 Introduction

Social media platforms have played an increasingly important role in capturing and communicating public opinion. Compared to traditional poll methods, where respondents are asked to fill out surveys on specific topics, social media allow people to share opinions on any topic and thus, give researchers insight into what the general public views as the salient topics of the day. However, researchers using social media do not always have the demographic features of those conversing on social media. To make this source of data comparable to surveys, it is important to understand the demographic characteristics of those using it.

There have been numerous studies on demographic inference for a range of demographics including age, gender, and location (Hinds and Joinson, 2018; Huang and Carley, 2019; Sakaki et al., 2014; Chamberlain et al., 2017; Liu et al., 2021; Al Zamal et al., 2012; Preoţiuc-Pietro and Ungar, 2018; Pennacchiotti and Popescu, 2011). However, because of the noiseness of the posts, the variability in the content users share, and the lack of dataset annotation for this problem, work still remains (Mneimneh et al., 2021).

Figure 1: An example of a multi-modal tweet. If we only consider the text "Yes, I'm Hiking! Young and Energetic!" we may make the wrong inference for age. The image, however, can provide important additional information.

In this paper, we study two important demographic attributes for social science research, gender and age. Most previous work focuses on developing a model using biography, tweet text, profile image, network, etc. However, as Liu and Singh point out, when analyzing popular hashtag streams, e.g. #blacklivesmatter or #maga, collecting biographies and user networks can be particularly costly. It is more common for researchers studying hashtags to have only post content, but want to answer the question - what are the demographics of those who are using this hashtag? Because this is the setting many computational social scientists work in, we emulate that setting by only using post text and post images as training data for our task.

While a number of studies use post text (Liu et al., 2021; Nguyen et al., 2013) and some use profile image (Vijayaraghavan et al., 2017; Sakaki et al., 2014; Wang et al., 2019), to the best of our knowledge, this is the first study that considers a combination of text and visual information from posts within a deep learning (DL) framework for demographic inference. Our decision to do this stems from two observations. First, according to a recent analysis of post content, more than 42% of tweets contain images (Gui et al., 2019), highlighting their

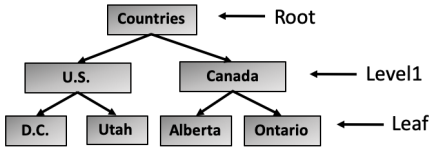

Figure 2: An example of hierarchical classification

prevalence on Twitter. Second, having both images and text can provide richer context when available. For example, Figure 1 shows an example of a multi-modal tweet. If we only consider the text "Yes, I'm Hiking! Young and Energetic!", our understanding of the age of the user who posted the tweet may be incomplete, leading to a wrong conclusion. Therefore, understanding the relationship between text and images may be beneficial to further improve the performance of demographic inference. We pause to mention that users share different levels of text and image data. Therefore, any classifier we build should be able to use either or both pieces of information effectively.

Some research has attempted to infer the relationship between the text and images of Twitter posts (Vempala and Preoţiuc-Pietro, 2019) or infer demographics using a user's posts images (Sakaki et al., 2014). However, because of the requirement for manual annotation of images and the focus on a single mode of information (text or images), we believe new multi-modal models are an important new direction.

Toward that end, we propose using contrastive learning for demographic inference. Contrastive learning is a self-supervised learning method that enables models to learn about data without data labels. Our approach is to use an existing neural model (CLIP) trained on a large set of [image,text] pairs (Radford et al., 2021) to learn visual concepts from images and language concepts from text and then use this information to improve demographic inference. CLIP maps text and images into the same embedding space. For our task, this is important since we may have a different amount of text and image data for different users. Therefore, even if some of our data does not contain text or images, CLIP is still able to encode text/images with both text and visual information from the embedding.

While our primary interest is on understanding if contrastive learning (learning from either or both text and image posts) is beneficial for demographic inference, we also have concerns about the lack of high quality labeled data for different demographic inference tasks. Most labeled sets are fairly small. Some research has investigated ways to label data at scale through Mechanical Turk and Wikidata (Vijayaraghavan et al., 2017; Liu et al., 2021; Sakaki et al., 2014; Chen et al., 2015). However, it is still unclear whether or not it is reasonable to combine or augment one dataset with other datasets for this task. We explore this notion by looking at how the inference quality of a small, labeled data changes when augmenting it with larger datasets that independently contain high quality examples for gender and age inference.

In summary, this paper makes the following contributions. (1) We combine knowledge from tweet text and tweet images within a neural network using an attention mechanism (Bahdanau et al., 2014) to improve the F1 score of binary gender and age inference by around 5% on average, respectively. (2) We propose the first model in demographic inference to use a contrastive learning approach for feature generation, thereby more explicitly exploring the relationship between text and images. (3) We incorporate hierarchical classification into our basic neural model to improve inference accuracy for multi-bin age by 1% to 2%. (4) We apply our model to a small dataset and analyze the value of augmenting the data with other datasets having high accuracy in their domains and show that this strategy improves the F1-score by 3% on average. (5) We make our IMDB dataset publicly available.

We pause to note that based on our data availability, we consider simplified versions of both gender and age. For gender, we consider the binary version of the task with male and female only because our ground truth data contains only those two classes. For age, we consider a binary task with two age bins and a multi-class version with three age bins.

2 Related Literature

It is a growing area of interest to predict attributes of social media as these platforms have become a place where people share their opinion on a range of issues. Most research on demographic inference uses classic algorithms such as logistic regression (LR), support vector machines (SVM), random forest (RF), usually with bag of words as features (Pennacchiotti and Popescu, 2011; Chen et al., 2015; Vijayaraghavan et al., 2017; Liu et al., 2021; Liu and Singh, 2021; Nguyen et al., 2013). Some papers have investigated using stylistic features, e.g.,

punctuation (Rosenthal and McKeown, 2011; Liu et al., 2021). In recent years, there has been more research using deep learning models. Miura and colleagues propose a model for location inference that combines tweet text, biography and network data using an attention mechanism (Miura et al., 2017). A graph-based Recursive Neural Networks (RNN) (Elman, 1990) using word embeddings is proposed by Mac Kim et al.. In their model, they make use of both user's posts and posts from user's network. Wood-Doughty and colleagues introduce a model that uses name alone with a Long-Short Term Memory (LSTM) network (Bengio et al., 1994) to infer gender (Wood-Doughty et al., 2018). Liu et al. use BERT to generate embeddings as model features. In their work, they also develop a fine-tuned BERT model, pretrained using Siamese network with SNLI (Bowman et al., 2015) and MNLI dataset (Williams et al., 2017) and achieve state of the art performance. Liu and Singh present a hierarchical model where they use GRU with an attention layer to separately train emoji component (using word embedding and convolutional neural network) and text component (using BERT), and then an attention layer is adopted to combine the two components. However, both (Liu and Singh, 2021; Liu et al., 2021) use BERT to process text only, missing potentially important context provided by visual attributes.

To explore the role images play on demographic inference, several studies propose using profile pictures. Scale invariant feature transformation (SIFT)(Lowe, 1999) is adopted by Chen et al. for image feature extraction. Vijayaraghavan and colleagues (Vijayaraghavan et al., 2017) extract feature representation from the profile images using the inception architecture (Szegedy et al., 2016), which is 48 layers deep and trained on more than a million images from the ImageNet database. Similar to Inception, ResNet50, a convolutional neural network that is 50 layers deep and trained on ImageNet, has also been applied in many computer vision tasks (He et al., 2021). But none of these studies use images from posts in demographic inference tasks. A model that combines text and images of posts is introduced by Sakaki et al.. In their work, they build an image classifier using SVM and post images with 10 labels, a categorization based on observation on a small-scale dataset by Ma et al.. Vempala and Preoţiuc-Pietro investigate the relationship between text and image of 4,471 tweets. But they did not apply the knowledge to downstream tasks. Morever, the data used in (Ma et al., 2014; Vempala and Preoţiuc-Pietro, 2019) is not large-scale and that makes the model less generalizable to transfer to another dataset.

Finally, previous works fail to consider how to achieve a comparable performance when text or images are not available. Thus, in this paper, we use CLIP, a pretrained model based on a contrastive approach using 400 millions of image and text pairs. To the best of our knowledge, this is the first time for contrastive learning to be applied for demographic inference.

There have been a few studies that investigated hierarchical classification on location inference. Mahmud et al. develop a two-level hierarchical location classifier that predicts a country location and then the city label within the former. Wing and Baldridge build a hierarchical tree as the classification structure. However, both methods have to train one classifier separately for many times, which is quite time-consuming and may encounter a situation where there is not enough data for one classifier. Huang and Carley propose a model that trains the country and city simultaneously and thus, greatly reduce the time and effort on training. To the best of our knowledge, there is no research that studies label hierarchies on age inference. In this paper, we adopt a similar approach as Huang and Carley. However, instead of fixing the values of the hyper-parameters like them, we also explore using automatic computation, i.e., let the model learn the appropriate hyper-parameters' values during training.

3 Experimental Design

3.1 Model Overview

Figure 3 shows the architecture of the proposed contrastive, multi-modal learning model (CMM). Tweet text and images extracted from posts are input into CLIP to get two separate embedding spaces so that the the text component and the image component can be trained independently. Specifically, they are each input into a RNN with attention. We use the gated recurrent unit (GRU) (Cho et al., 2014), a variant of the RNN that can avoid the problem of vanishing and exploding gradients(Cho et al., 2014). For age with 3 bins, following the attention layer, a hierarchical classification learning process occurs where constraints learned from coarse-grained predictions are input to fine-grained

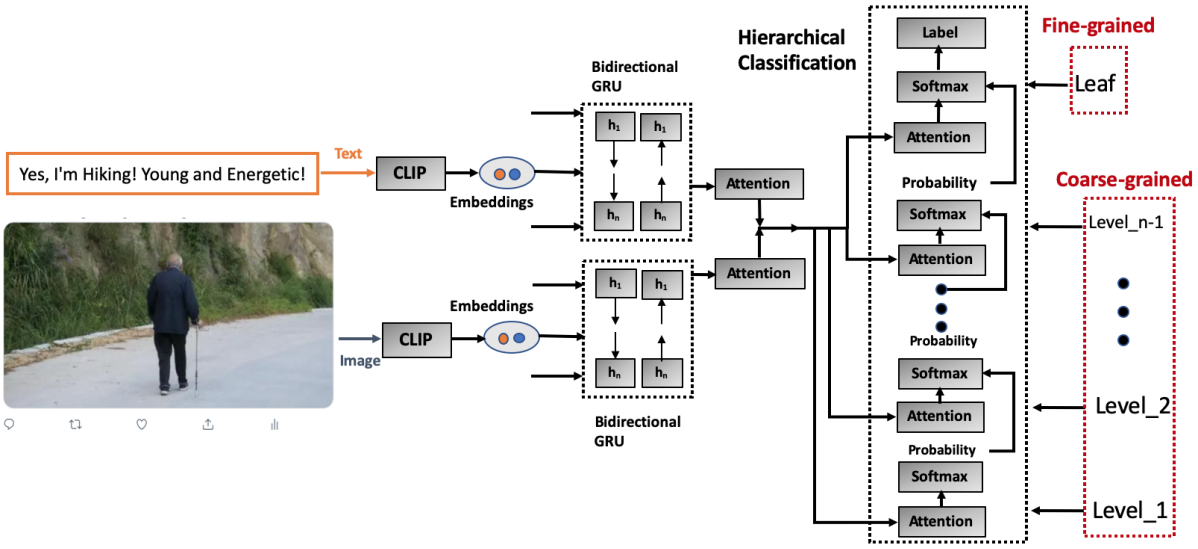

Figure 3: Overview of the proposed model

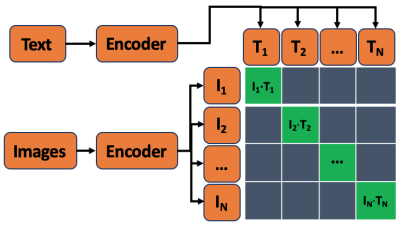

Figure 4: Contrastive Pretraining

predictions. The final prediction is at the leaf level of the hierarchy. For gender and age with two bins, since they are binary classification, following the attention layer, the output will move into the leaf nodes in the hierarchy directly without any coarse-grained predictions. The remainder of this section describes the details of our model.

3.2 CLIP

CLIP, developed by OpenAI, is a model that attempts to connect text and images. It is trained on a set of 400 million (image I_j, text T_j) pairs and has a contrastive objective as shown in figure 4 (Radford et al., 2021). The model tries to learn the relationship between an entire sentence and the image it describes with the goal of maximizing the similarity of diagonal (the green area - $(I_1T_1, I_2T_2, ..., I_NT_N)$ and minimize the remaining area. Research has shown that CLIP is capable of (1) generating captions given an image as it can "understand" the objects of an image, (2) helping generate an image based on text, (3) predicting the most relevant text snippet given an image, (4) conducting the image classification task in computer vision with zero-shot capabilities (Radford et al., 2021; Galatolo et al., 2021; Patashnik et al., 2021). In this paper, we show that CLIP is not only limited to the tasks above. Rather, due to the large amount of (text, image) pairs on Twitter posts, CLIP can also be applied to the classification task of demographic inference of Twitter users.

3.3 Image

Since images are widely used on Twitter, we hypothesis that mapping images into an embedding space using CLIP will provide additional contextual cues that improve demographic inference. Specifically, we extract image features from the CLIP model directly and generate the embedding vector with a dimension D to represent an image.

Image Encoder: For each user, we have a set of n images ordered based on the date/time of the post. Given the sorted post images and their vector representation g, we use a bidirectional GRU to encode tweet images and get the representation h.

$$\overrightarrow{h_i} = \overrightarrow{GRU}(g_i), i \subset [1, N]$$

$$\overleftarrow{h_i} = \overleftarrow{GRU}(g_i), i \subset [N, 1]$$

Next, we concatenate $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ to get an annotation of the post image i, i.e., $\boldsymbol{h}_i = [\overrightarrow{h_i}, \overleftarrow{h_i}]$. Here, $\boldsymbol{h}_i$ summarizes the post image near image i.

Image Attention: An attention mechanism is used to reward tweet images that help correctly classifying a user. This yields

$$\boldsymbol{u}_i = tanh(\boldsymbol{W}_s h_i + \boldsymbol{b}_i),$$

$$\boldsymbol{\alpha}_i = \frac{exp(\boldsymbol{u}_i^T c_i)}{\sum_i exp(\boldsymbol{c}_i^T \boldsymbol{c}_i)},$$

$$\boldsymbol{v}_i = \sum_i \alpha_i \boldsymbol{h}_i$$

With image annotation $\boldsymbol{h}_i$ fed into a MLP, we obtain the hidden representation $\boldsymbol{u}_i$. Then we use a context vector $\boldsymbol{c}_i$, which lets us measure the importance of each tweet image and get a normalized weight $\boldsymbol{\alpha}_i$ through a softmax function. Finally, with the weights computed, we get the image vector of a user $\boldsymbol{v}_i$ as a weighted sum of the image annotations. $\boldsymbol{W}, \boldsymbol{b}$ are the parameter matrices and bias.

3.4 Text

For the text representation, we use the CLIP model directly to generate the embeddings with a dimension D, having tweet text as input. Similar to image encoder, we use a GRU structure to encode all the tweet text for each user. Next, we adopt an attention mechanism so that the model is able to selectively focus on valuable parts of the input text for our task and learn the association between them.

3.5 Feature Fusion

Similar to previous work in this area (Liu and Singh, 2021), we combine different components using attention.[1] Specifically, our text and image components are trained independently with different weights and biases. Then we input each of them into the attention layer to combine and summarize the image and text and obtain the representation vector $\boldsymbol{v}$.

3.6 Classification Layer

This section begins by explaining our approach to hierarchical classification for multi-class inference, age in our case. We then explain the simplified model for the binary case.

3.6.1 Hierarchical Age Prediction

Suppose age is represented as a set of k bins, $(x_0, x_1, ...x_k)$, where x_1 contains the youngest group, and x_k contains the oldest. In the case where $k > 2$, we are interested in building a hierarchical tree structure that can be used to incorporate additional constraints from coarser bins into the learning process. For example, if we want to determine if a person is between the ages of 18 and 30, it can be useful to know that he/she fits into a coarser category, e.g. having an age between 18 and 45. Next, with x being the leaf node (level n), we build the parent level bins by merging 2 continuous bins into 1, i.e., integrating $(x_0, x_1), (x_2, x_3), ..., (x_{k-1}, x_k)$ and we get $round(k/2)$ bins. After iterating though all bins from level n, we get the bins for $n-1$ level. We repeat the same process recursively until there are two bins left, which makes it a binary classification.

Typically, $k = 3$, meaning that we have a three level hierarchy with a root node (level 0), an internal level (level 1) with two bins and the leaf level with three bins (level 2). The probability for the coarse-grained age (level 1) is computed by a softmax function

$$\boldsymbol{P}_c = softmax(\boldsymbol{W}_c \boldsymbol{v}_c + \boldsymbol{b}_c)$$

where $W_c \subset R^{N_c * D}$is a linear projection parameter, $b_c \subset R^{N_c}$ is a bias term, and N_c is the number of bins, and v_c is the vector representation from the image and text component obtained from the attention mechanism. After getting the probability for coarse-grained age bin, we use it to constrain the fine-grained age prediction(level 2)

$$\boldsymbol{P}_f = softmax(\boldsymbol{W}_f \boldsymbol{v}_f + \boldsymbol{b}_f + \lambda \boldsymbol{P}_c Bias)$$

Where $\boldsymbol{W}_f \subset \boldsymbol{R}^{N_f * D}$ is a linear projection parameter, $\boldsymbol{b}_f \subset \boldsymbol{R}^{N_f}$ is a bias term, $\boldsymbol{N}_f$ is the number of fine-grained age bins and $\boldsymbol{v}_f$ is the vector representation. $Bias \subset \boldsymbol{R}^{N_c * N_f}$ is the coarse-grained to fine-grained correlation matrix. If fine-grained age j belongs to coarse-grained age i, then $Bias_{ij}$ is 0, otherwise -1. But the value of this hyper-parameter can be tuned. λ is a penalty term. The larger of λ, the stronger of the coarse-grained age constraint.

We minimize the sum of two cross-entropy losses for coarse-level prediction and fine-level prediction.

$$loss = -(\boldsymbol{Y}_f * log \boldsymbol{p}_f + \alpha \boldsymbol{Y}_c * log \boldsymbol{P}_c)$$

where $\boldsymbol{Y}_f$ and $\boldsymbol{Y}_c$ are one-hot encodings of fine-grained and coarse-grained labels. α is the weight to control the importance of coarse-grained age supervision signal. We experimented with two settings: (1) Fix the value of the matrix $Bias$ and λ. (2) Let the model learn the correlation matrix and λ during training, making the process fully automatic.

The process for age with two bins and gender is similar to the fine-grained prediction for age with multiple-label except the term $\lambda \boldsymbol{P_c} Bias$ that

[1] Based on our experiments, CLIP is able to encode emojis. So, we do not incorporate emojis in our model.

Demographics		Category	Count			
			Wiki	IMDB	MI	Combined
Age	Bin 2	<45	7538	1898	324	787
		>=45	3731	1467	348	721
	Bin 3	<35	5206	807	178	465
		35-54	3907	2013	296	592
		>=55	2156	545	198	451
Gender	-	Female	3335	1454	289	720
	-	Male	7891	1911	383	788

Table 1: Ground truth data distribution

is introduced by coarse age prediction is removed. Specifically, After the feature fusion layer, we send the data into a fully connected (FC) layer and Softmax layer and get the final prediction.

4 Empirical Evaluation

For this evaluation, we use the following four datasets.

Wiki Similar to Liu et al., we use a dataset constructed from Wikidata (Vrandečić and Krötzsch, 2014) that contains a mapping between user demographics and Twitter handles. We then use the Twitter API to retrieve users' most recent posts and the image link within a post. We then download the images using the links.

IMDB Beginning with a public dataset containing the demographic information of actors and actresses in IMDB, we used different celebrity lists to scrape the Twitter handles of different celebrities in 2017. We then used the Twitter API to collect tweets of each celebrity.[2] There are no overlapping handles between the Wiki data and the IMDB data.

Survey Data Our research team conducted a nationally representative survey that asked respondents about a number of opinions. Those respondents who used Twitter were also asked if they would consent to allow our research team to download their tweets. This dataset contains tweets and images from those who consented.[3]

Combined data For this dataset, we combine our different sources. We use the entire Survey dataset, and randomly sample similar numbers of users from Wiki dataset and IMDB dataset.

We use the following pre-processing procedure: 1) remove users that have less than 20 English tweets, 2) remove users in the Wiki dataset that do not have at least one post image (this is done to show the impact of images on performance), 3) remove stopwords, handles, and mentions for the classic models, and lowercase all of the words for the classic models.

Table 1 shows the number of users in each dataset for gender and age category. For age, 45 defines a new era of adulthood based on the Levinson adult development model (Levinson, 1986). Thus, we choose 45 as the 2-bin dividing line. The 3-bin boundaries were identified by social science experts. When necessary, we randomly sample from the group using the Python library *imblearn* (Lemaî et al., 2017) in order to create more balanced datasets.

4.1 Baseline Models

For text analysis, the classic models we compare to are logistic regression (LR) (Nguyen et al., 2013), SVMs (Chen et al., 2015), and Random Forest (Cornelisse and Pillai, 2020; Vijayaraghavan et al., 2017). The neural network models we compare to are Vanilla BERT (Liu et al., 2021), Siamese BERT (Liu et al., 2021), the hierarchical text/emoji model (Liu and Singh, 2021), and ResNet-50 and Inception-v3 for images.

4.2 Experiment settings

We use 2 NVIDIA Tesla P4 GPUs each having 16 GBs of memory. We use the Adam update rule (Kingma and Ba, 2014) to optimize our model. Weight, bias and context vector are randomly initialize for the attention layers and then normalized with a mean value of 0 and a standard deviation of 0.05. They are jointly learned during training. Gradients are clipped between -1 and 1. We also tune hyperparameters for model optimization.[4]

We use 5-fold cross validation with a training and validation set. We also have a separate holdout set. We do the 5-fold cross validation three times with three different random seeds. We report the averaged F1 score with confidence intervals (appendix) and the holdout set F1 score.

4.3 Experiment results

Table 2 presents a comparison between previous models and our models using the Wiki dataset and the Survey dataset. We present three training data variants: text only, images only, and a combination of text and images. Table 3 shows the comparison between previously proposed models and our

[2]The preprocessed data can be found at *removed for review*.

[3]Details about the project and the IRB have been removed for review.

[4]Batch size for Wiki and IMDB is 32. For Survey and the combined dataset, we use 16. BERT embedding dimension is 768 and CLIP is 512. Both initial values of λ and α are set as 1 and learning rate is set as 0.0001

Model	Wiki			Survery		
	Bin 2	Bin 3	Gender	Bin 2	Bin 3	Gender
Text						
Unigram-RF	0.821	0.653	0.839	0.686	0.535	0.687
Nguyen et al.	0.777	0.636	0.802	0.662	0.497	0.674
Chen et al.	0.809	0.645	0.813	0.646	0.496	0.65
Vanilla BERT	0.784	0.605	0.869	0.684	0.55	0.714
Siamese BERT	0.790	0.610	0.871	0.672	0.531	0.721
Liu et al.	0.838	0.671	0.876	0.718	0.578	0.696
CMM	**0.855**	**0.728**	**0.889**	0.739	0.651	0.745
Image						
Inception	0.770	0.587	0.832	0.526	0.325	0.512
Resnet50	0.762	0.586	0.831	0.603	0.367	0.481
CMM	**0.851**	**0.723**	**0.917**	**0.692**	**0.531**	**0.698**
Text + Image						
CMM	0.861	0.742	0.945	0.713	0.577	0.702
HCMM	-	0.754	-		**0.629**	
AHCMM	-	0.756	-		0.624	

Table 2: Age and Gender result for Wiki dataset. - means non-applicable and Bin # refers to the number of bins for age. Red indicates the highest performance and bold corresponds to the highest within a group

Model	IMDB			Combined		
	Bin 2	**Bin 3**	**Gender**	**Bin 2**	**Bin 3**	**Gender**
Unigram-RF	0.716	0.585	0.824	0.73	0.578	0.747
Nguyen et al.	0.71	0.573	0.826	0.719	0.508	0.719
Chen et al.	0.724	0.572	0.826	0.766	0.534	0.755
Vanilla BERT	0.692	0.553	0.825	0.731	0.554	0.759
Siamese BERT	0.669	0.526	0.834	0.724	0.535	0.775
Liu et al.	0.713	0.581	0.839	0.757	0.558	0.775
CMM	**0.749**	0.608	**0.873**	**0.782**	**0.633**	**0.791**
HCMM	-	**0.628**	-	-	0.627	-
AHCMM	-	0.625	-	-	0.611	-

Table 3: Age and Gender result for the combined dataset. - means non-applicable and Bin # refers to the number of bins for age

models for the IMDB dataset and the combined dataset.

Wiki analysis: We can see that for the binary classifiers, CMM using the combined text and image training dataset performs better than the state of the art classic and neural models by approximately 2% to 9% for age and 6% to 11% for gender. What is also interesting is that when using only a single mode of data, i.e. either text or images, CMM still performs better than the current state of the art. It is interesting to note that using text alone or images alone result in comparable age F1 scores for CMM. On the other hand, for gender, CMM using images for training data performs much better than text.

Finally, for the three bin age classification, both HCMM and AHCMM perform 8% to 17% better than the state of the art, 1% to 2% better than CMM, highlighting the value of both the contextual and hierarchical components of our model. The similarity in F1 score between AHCMM and HCMM suggests that automated learning does not help much on this task when compared to fixing hyper-parameters. Similar to our 2-bin age results, using only text or only images for training data results in similar F1 scores for CMM.

Overall, the F1 score of the proposed model is 2.3%, 8.5%, 6.9% higher than the best previous model for age with 2 bins, age with 3 bins and gender, respectively, indicating that contrastive learning is useful for demographic inference.

Survey analysis: Similar to the Wiki data, our models perform better than the state of the art and HCMM and AHCMM perform better than CMM. While the model built using the text and image training data performs well, the model using only text training data is the best, between 2% and 7% better than the state of the art. A closer examination of the proportion of text posts and images shows that there are significantly more text posts than images for both Wiki data and Survey data (see Appendix). However, given that the overall amount of training data is much less for the Survey dataset, we surmise that in a more constrained environment, having text can be more beneficial for demographic inference than having images.

We pause to note how much worse the survey models perform compared to the models using the Wiki data. That is a direct result of the small amount of training data and the reason we consider the combined data.

IMDB analysis: Recall that for the IMDB dataset, we only have text posts. This dataset gives us the opportunity to focus on the value of CLIP given a single type of data.

For this case, the classic models perform similarly. The neural models are generally not as good as the classic models. Our proposed model achieves the highest F1 score for binary age, 2.5% higher than the best previous model. For age with 3 bins, we again see both HCMM and AHCMM are around 2% to 3% higher than CMM, with HCMM and AHCMM having marginal difference with each other. The best proposed model for age with 3 bins is 4.3% higher than the best previous model.

Finally, for gender, we see that the neural models are better than the classic models and that our model has an F1 score that is 3.4% higher than the best state of the art model.

Combined analysis

Similar to our previous findings, CMM performs better than the state of the art for age and gender, ranging from 1.6% better for 2-bin age and gender to over 6% for 3-bin age.

However, what is more important is whether or

Dataset	Bin 2		Bin 3		Gender	
	Sep	SC	Sep	SC	Sep	SC
Survey	0.739±0.022	0.747±0.03	0.649±0.027	0.64±0.052	0.725±0.068	0.745±0.017
Wiki	0.791±0.022	0.77±0.023	0.624±0.02	0.613±0.025	0.772±0.036	0.812±0.026
IMDB	0.786±0.033	0.801±0.03	0.569±0.03	0.634±0.024	0.826±0.028	0.85±0.024
Combined	0.775±0.013		0.641±0.014		0.807±0.013	

Table 4: Result for the 3 separate datasets and combined data

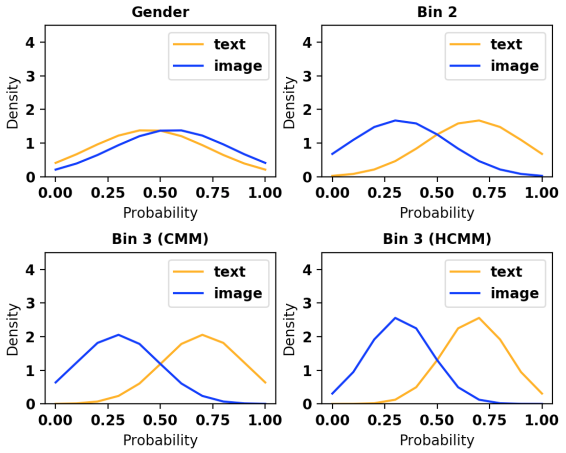

Figure 5: : Estimated probability density functions of the attention layer

not combining data leads to better results than using a smaller survey data set by itself. We find that combining the datasets leads to better F1 scores than the survey model on its own.

Table 4 broadens our data augmentation comparison. It shows separate F1 scores (Sep) when the 3 datasets (the survey data and the sampled data from Wiki and IMDB) are trained independently, F1 scores for each separate dataset within the combined dataset (SC) trained as a whole, and the overall F1 score for the combined dataset.

The table highlights the following. The combined dataset has a smaller confidence interval than the 3 separate datasets. The survey data generally has improved performance with data combined. The F1 score for the Wiki dataset is lowered. The performance of IMDB improves in all cases. The takeaway message is that if a researcher has a set of small datasets, it can be beneficial to augment them. Although intuitively we thought Wiki and IMDB dataset were more similar since both of them are mainly celebrities, the result suggests that the two datasets do not supplement each other.

4.4 Analysis of Attention

In the evaluation, the proposed model for Wiki dataset has shown effectiveness at unifying text and image representations through F1 score. However, details of the unification processes are not clear from the model outputs. To gain insight into the unification processes, we analyze the states of the final attention layer that combines the image component and text component. Figure 5 shows the kernel density estimation for different gender, age with 2 bins, 3 bins using CMM and HCMM.[5] We see that for gender, text and images have similar probabilities from the model with images being slightly more important. In general, both text and images provide valuable information. For age, it is apparent that the model assigns higher probabilities to text, highlighting its higher value.

5 Conclusions and Future Work

This paper proposes using contrastive learning as a way to incorporate knowledge from text and/or images for demographic inference, thereby more explicitly exploring the relationship between text and images. We combine knowledge from tweet text and tweet images within a neural network to improve the F1 score of gender and age inference. We also incorporate hierarchical classification into our basic neural model to further improve inference accuracy. For future work, we would like to pretrain CLIP with Twitter data as it may provide additional information to improve the performance.

6 Ethical Considerations

We acknowledge that demographics prediction can have ethical implications. While automated models could provide valuable information on understanding people's opinion, errors occur that may lead to possible equity and justice related consequences.We also believe that privacy expectations should not be compromised. For this reason we use publicly available Wikidata and IMDB data (users publicly share their handles) and Survey data, (users agree to share their information for research purpose only).

[5]Here we did not show the visualization result using AHCMM model due to space limit. The HCMM and AHCMM results are very similar.

References

F. Al Zamal, W. Liu, and D. Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *AAAI Conference on Weblogs and Social Media*.

D. Bahdanau, K. Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Y. Bengio, P. Simard, and P. Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*.

S. Bowman, A., C. Potts, and C. Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

B. Chamberlain, C. Humby, and M. Deisenroth. 2017. Probabilistic inference of twitter users' age based on what they follow. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.

X. Chen, Y. Wang, E. Agichtein, and F. Wang. 2015. A comparative study of demographic attribute inference in twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*.

K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Joran Cornelisse and Reshmi Gopalakrishna Pillai. 2020. Age inference on twitter using sage and tf-igm. In *Proceedings of the International Conference on Natural Language Processing and Information Retrieval*.

J. Elman. 1990. Finding structure in time. *Cognitive science*.

F. Galatolo, M. Cimino, and G. Vaglini. 2021. Generating images from caption and vice versa via clip-guided generative latent space search. *arXiv preprint arXiv:2102.01645*.

T. Gui, L. Zhu, Q. Zhang, M. Peng, X. Zhou, K. Ding, and Z. Chen. 2019. Cooperative multimodal approach to depression detection in twitter. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

K. He, X. Zhang, S. Ren, and J. Sun. 2021. Deep residual learning for image recognition. 2015. *arXiv preprint arXiv:1512.03385*.

J. Hinds and A. Joinson. 2018. What demographic attributes do our digital footprints reveal? a systematic review. *PloS one*.

B. Huang and K. Carley. 2019. A hierarchical location prediction neural network for twitter user geolocation. *arXiv preprint arXiv:1910.12941*.

D. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

G. Lemaî, F. Nogueira, and C. Aridas. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*.

D. Levinson. 1986. A conception of adult development. *American psychologist*.

Y. Liu and L. Singh. 2021. Age inference using a hierarchical attention neural network. In *The Conference on Information and Knowledge Management*.

Y. Liu, L. Singh, and Z. Mneimneh. 2021. A comparative analysis of classic and deep learning models for inferring gender and age of twitter users. In *Proceedings of the International Conference on Deep Learning Theory and Applications*.

D. Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the IEEE international conference on computer vision*.

X. Ma, Y. Tsuboshita, and N. Kato. 2014. Gender estimation for sns user profiling using automatic image annotation. In *IEEE International Conference on Multimedia and Expo Workshops*.

S. Mac Kim, Q. Xu, L. Qu, S. Wan, and C. Paris. 2017. Demographic inference on twitter using recursive neural networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

J. Mahmud, J. Nichols, and C. Drews. 2012. Where is this tweet from? inferring home locations of twitter users. In *International AAAI Conference on Weblogs and Social Media*.

Y. Miura, M. Taniguchi, T. Taniguchi, and T. Ohkuma. 2017. Unifying text, metadata, and user network representations with a neural network for geolocation prediction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Z. Mneimneh, J. Pasek, L. Singh, R. Best, L. Bode, E. Bruch, C. Budak, P. Davis-Kean, K. Donato, N. Ellison, et al. 2021. Data acquisition, sampling, and data preparation considerations for quantitative social science research using social media data.

D. Nguyen, R. Gravel, and T. Trieschnigg, D.and Meder. 2013. " how old do you think i am?" a study of language and age in twitter. In *AAAI Conference on Weblogs and Social Media*.

O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

M. Pennacchiotti and A. Popescu. 2011. A machine learning approach to twitter user classification. In *AAAI Conference on Weblogs and Social Media*.

D. Preoţiuc-Pietro and L. Ungar. 2018. User-level race and ethnicity predictors from twitter text. In *Conference on Computational Linguistics*.

A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

S. Rosenthal and K. McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Association for Computational Linguistics: Human Language Technologies*.

S. Sakaki, Y. Miura, X. Ma, K. Hattori, and T. Ohkuma. 2014. Twitter user gender inference using combined analysis of text and image processing. In *Workshop on Vision and Language*.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

A. Vempala and D. Preoţiuc-Pietro. 2019. Categorizing and inferring the relationship between the text and image of twitter posts. In *Proceedings of the annual meeting of the Association for Computational Linguistics*.

P. Vijayaraghavan, S. Vosoughi, and D. Roy. 2017. Twitter demographic classification using deep multimodal multi-task learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

D. Vrandečić and M. Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*.

Z. Wang, S. Hale, D. Adelani, P. Grabowicz, T. Hartman, F. Flöck, and D. Jurgens. 2019. Demographic inference and representative population estimates from multilingual social media data. In *The world wide web conference*.

A. Williams, N. Nangia, and S. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

B. Wing and J. Baldridge. 2014. Hierarchical discriminative classification for text-based geolocation. In *Proceedings of conference on empirical methods in natural language processing*.

Z. Wood-Doughty, N. Andrews, R. Marvin, and M. Dredze. 2018. Predicting twitter user demographics from names alone. In *Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*.

	Wiki		Survey		IMDB
	Text	Image	Text	Image	Text
Total	1821323	473890	71728	24720	744162
Mean	127	42	149	56	193
Median	200	26	200	19	200

Table 5: Statistics over the Wiki, IMDB and Survey data

Model	Data Type	Wiki		
		Bin 2	**Bin 3**	**Gender**
CMM	Text	0.858+0.005	0.73+0.008	0.889+0.006
CMM	Image	0.853+0.007	0.73+0.01	0.917+0.004
CMM	Text&Image	0.87+0.007	0.746+0.01	0.935+0.006
HCMM	Text&Image	-	0.756+0.008	-
AHCMM	Text&Image	-	0.754+0.008	-

Table 6: Mean and (0.95) confidence interval F1 score for Wiki dataset

A Appendix

Table 5 demonstrates the statistics information of tweet text and images. Table 6 shows the result for Wiki dataset. Table 7 shows the detailed F1 score for IMDB dataset and Combined dataset.

Model	IMDB			Combined		
	Bin 2	**Bin 3**	**Gender**	**Bin 2**	**Bin 3**	**Gender**
CMM	0.748±0.01	0.612±0.019	0.868±0.008	0.775±0.013	0.64±0.014	0.807±0.013
HCMM	-	0.624±0.012	-		0.628±0.017	-
AHCMM	-	0.623±0.019	-		0.63±0.019	-

Table 7: Mean and (0.95) confidence interval F1 score for IMDB and the combined dataset