Mathematical Biosciences and Engineering

MBE, 20(9): 15962–15981.
DOI: 10.3934/mbe.2023711
Received: 17 October 2022
Revised: 27 June 2023
Accepted: 20 July 2023
Published: 03 August 2023

http://www.aimspress.com/journal/MBE

Research article

Twitter-based gender recognition using transformers

Zahra Movahedi Nia[1,2], Ali Ahmadi[3,4], Bruce Mellado[1,5], Jianhong Wu[1,2], James Orbinski[1,6], Ali Asgary[1,4] and Jude D. Kong[1,2,*]

[1] Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC), York University, Canada
[2] Laboratory for Industrial and Applied Mathematics, York University, Canada
[3] K.N Toosi University, Faculty of Computer Engineering, Tehran, Iran
[4] Advanced Disaster, Emergency and Rapid-Response Simulation (ADERSIM), York University, Toronto, Ontario, Canada
[5] School of Physics, Institute for Collider Particle Physics, University of Witwatersrand, Johannesburg, South Africa
[6] Dahdaleh Institute for Global Health Research, York University, Canada

* **Correspondence:** Email: jdkong@yorku.ca.

Abstract: Social media contains useful information about people and society that could help advance research in many different areas of health (e.g. by applying opinion mining, emotion/sentiment analysis and statistical analysis) such as mental health, health surveillance, socio-economic inequality and gender vulnerability. User demographics provide rich information that could help study the subject further. However, user demographics such as gender are considered private and are not freely available. In this study, we propose a model based on transformers to predict the user's gender from their images and tweets. The image-based classification model is trained in two different methods: using the profile image of the user and using various image contents posted by the user on Twitter. For the first method a Twitter gender recognition dataset, publicly available on Kaggle and for the second method the PAN-18 dataset is used. Several transformer models, i.e. vision transformers (ViT), LeViT and Swin Transformer are fine-tuned for both of the image datasets and then compared. Next, different transformer models, namely, bidirectional encoders representations from transformers (BERT), RoBERTa and ELECTRA are fine-tuned to recognize the user's gender by their tweets. This is highly beneficial, because not all users provide an image that indicates their gender. The gender of such users could be detected from their tweets. The significance of the image and text classification models were evaluated using the Mann-Whitney U test. Finally, the combination model improved the accuracy of image and text classification models by 11.73 and 5.26% for the Kaggle dataset and by 8.55 and 9.8% for the PAN-18 dataset, respectively. This shows that the image and text classification models are capable of complementing each other by providing additional information to one another. Our overall multimodal method has an accuracy of 88.11% for the Kaggle and 89.24% for the PAN-18 dataset and outperforms state-of-the-art models. Our work benefits research that critically require user demographic information such as gender to further analyze and study social media content for health-related issues.

Keywords: BERT; ELECTRA; gender recognition; LeViT; RoBERTa; social media; Swin transformer; transformers; ViT

1. Introduction

People are progressively becoming active in social media, sharing their thoughts, beliefs, concerns and experiences. Consequently, a huge amount of useful information is produced that can help solve many problems in health such as mental health [1], health surveillance [2], public safety and policy [3,4], healthcare [5,6] and gender vulnerability [7,8]. User demographics provide social media-based research with essential information that can help study the issue from diverse perspectives. However, on most social media platforms, user information such as gender is considered private and therefore not freely available.

The COVID-19 pandemic has exacerbated global socio-economic inequalities, revealing how crises affect people differently according to their gender in troubling patterns which do not bode well for future resilience. Integrating governance at widening levels and mitigating the limited economic options of women are two examples of systematic challenges which require attention for human futurity. However, in many cases, even the data required to document and understand these challenges is not available. This paper addresses these systematic imperatives by providing a model for extracting users' gender on social media and helping researchers identify the elements of promising emergent governance frameworks to address local and global-scale socio-ecological challenges that disproportionately impact women.

Although many previous studies have focused on finding user information such as gender from text data [9–13], very few of them have considered using images. Combining image and text classification methods for finding users' genders can significantly increase the classification accuracy [14,15]. In this paper, we propose a multimodal approach to find social-media users' gender by combining text and image processing and adapting transformers.

Transformers are novel deep learning models that use a self-attention mechanism to identify and learn significant parts of a content [16]. The attention mechanism is a technique that is capable of enhancing and highlighting important parts of the content while downgrading other parts [17]. Self-attention is an attention mechanism that finds important tokens and their relations by comparing content with itself [18]. A token is usually a single word in natural language processing (NLP) and a group of pixels, known as a patch, that are processed together in computer vision. Since transformers can process tokens sequentially, they are suitable for both text and image processing [19,20].

Transformers were initially used for NLP and later on for computer vision. Before transformers, recurrent neural network (RNN) models such as long-short term memory (LSTM) and gated recurrent units (GRUs) with added attention layers on top of them were commonly used for NLP and convolutional neural networks (CNN) were dominantly used for vision. In 2017, transformers were introduced by keeping the attention layer and dropping the RNN part to speed up the training process for NLP. Recently, transformers have been used and performed very well in image recognition. BERT [21] and ViT [22] are one of the first models built using transformers and trained for text and image classification, respectively.

BERT, which has become very popular for NLP lately, was first developed in 2018 to improve GPT by looking at sequences of texts in a bidirectional way. GPT is a transformer-based model that was proposed by OpenAI in 2018, trained in an unsupervised manner and then fine-tuned for a specific supervised NLP task [23]. GPT includes 12-layers of transformer decoders with masked self-attention. For unsupervised learning, the model was pretrained for next-token prediction using an unpublished book dataset. Then the model was fine-tuned through labeled datasets for procedures such as classification, textual entailment, and sentiment analysis. This training technique is extremely favorable to NLP developers, since it performs very well when less labelled data is available.

BERT was presented in two different modes, BERT$_{\text{BASE}}$ and BERT$_{\text{LARGE}}$ which respectively include twelve layers of transformers with twelve-headed bidirectional self-attention and twenty-four layers of transformers with sixteen-headed bidirectional self-attention. Both models have been trained in an unsupervised manner for language modelling and next-sentence prediction, using a large corpus gathered from books and Wikipedia pages. This time consuming computationally-expensive pre-training phase resulted in learning contextual embeddings for tokens i.e., words, by BERT. BERT can then be fine-tuned to perform different NLP tasks such as question answering and language understanding, in a supervised manner.

Soon after, other models were developed to improve BERT. RoBERTa trains the BERT model with different hyperparameters, longer sequences and a larger batch size. Moreover, it applies dynamic masking for masked language modeling (MLM) rather than static masking which is usen in BERT and achieves significantly better results on different datasets [24]. XLNet replaces the autoencoding model of BERT with an autoregressive model and gains better results compared to BERT and RoBERTa [25]. ELECTRA substitutes the MLM pretraining method used in BERT with a replaced token detection method and outperforms the previous models in terms of accuracy while having less computational complexity [26].

After NLP, transformers were adjusted for constructing vision models using sequences of pixels/patches. Image GPT (iGPT) and ViT were the first vision models built with transformers. iGPT was developed in 2020 by OpenAI and trained in three different sizes, iGPT-S, iGPT-M and iGPT-L, which included 76 million, 455 million and 1.4 billion parameters, respectively. Since finding the relation between pixels is prohibitively complex in terms of memory and computation, iGPT reduces the resolution and color space of an image and then applies generative training on sequences of pixels using transformers [27]. ViT was developed in 2020 and published in 2021 by researchers from Google's Brain Team [28]. To decrease memory and computation complexities, ViT divides an image into 16 × 16 pixel sections for processing. Thus, a token is a 16 × 16 pixel piece of an image in ViT. Next, a learnable embedding vector is assigned to each token and along with positional embeddings are fed into a transformer architecture. Three different models are defined and trained for ViT, namely, ViT-Base, ViT-Large and ViT-Huge, which respectively, include twelve layers of transformers with twelve-head self-attention, twenty-four layers of transformers with sixteen-head self-attention and thirty-two layers of transformers with sixteen-head self-attention. The models have been pre-trained for image classification on different datasets including ImageNet, ImageNet-21k and JFT-300M and have had up to 99.74% accuracy. The authors found that when trained on large datasets (14–300 million images), ViT outperforms CNN-based models such as ResNet [29] and EfficientNet [30].

Afterwards, different vision models were proposed and built on top of ViT for image classification. DieT [31] was the first work to successfully train transformer-based models using mid-sized datasets (i.e. 1.2 million samples of ImageNet rather than 300 million images of JFT). A CNN was used as a teacher model for DieT to train the useful representations of input images. Hard and soft labeling were explored for this distillation approach, where the hard distillation was found to perform fairly better. Swin transformers [32] proposed hierarchical feature maps through merging image patches. It performs local attention using window partitioning, and uses shifted window approach to find cross-window connections. Several works have suggested to augment ViT with CNN architecture [33–35]. Convolutional vision transformer (CvT) [36] introduces CNNs to ViT to capture spatial structures and low-level details of image patches. CvT has a hierarchical design in which the sequence length progressively decreases while the token width increases. LeViT [37] used CNNs for image processing and feature extraction and passed the outcome as an input to a hierarchical ViT architecture. ViT has also been adjusted for carrying other image processing tasks such as object detection [38,39], segmentation [40] and image generation [41].

Previously, some works have used only the profile images to predicted user genders [42], while others have gathered several images posted by the users in social media to discover their gender [43,44]. In this work, we use transformer models to explore both of the methods and compare them in terms of accuracy. We use a gender classifier dataset available on Kaggle [45] and the PAN-2018 dataset [46,47] to build gender classification models based on profile images and image content posted by the user on social media, respectively. We fine-tune three vision models, i.e., ViT, LeViT and Swin Transformer to predict gender based on Twitter profile images (the Kaggle dataset) and ten different images posted by a user on Twitter (the PAN-18 dataset), respectively. In addition, we fine-tune three NLP models, i.e., BERT, RoBERTa and ELECTRA for text-based gender recognition using approximately 100 tweets posted by the user for both of the Kaggle and PAN-18 datasets. We found that concatenating several tweets improves the accuracy of the text-classification model. Likewise, concatenating several images posted on Twitter improves the accuracy of the image-classification model. Eventually, we combined the image- and text-classification models and found a high accuracy of 88.11 and 89.24% using transformers for the Kaggle and PAN-18 datasets, respectively. Our contribution to this work is threefold:

- We have fine-tuned and compared different state of the art transformer-based vision and text models for classification and evaluated their statistical significance using Mann-Whitney U test.
- We have completed the publicly available dataset on Kaggle and provided approximately 100 tweet ids for each female, male and brand classes. Therefore, we provide a great dataset that future works could build up on. Our dataset is publicly available at [48].
- Our work is extendable to other social media platforms such as Facebook and Reddit. This work paves the path for other research that require gender information of social media users for studying health-related issues.

We compared our model with state-of-the-art models and found that our multimodal method is superior to other methods in terms of accuracy.

In the following, Section 2 includes the literature review. Sections 3 and 4 present our proposed method and numerical results, respectively. A discussion is provided in Section 5, followed by conclusion and future work in Section 6.

2. Literature review

Finding gender from text has been practiced using different approaches [9–13]. Vashisth and Meehan [9] used different NLP methods for gender detection using Tweets, including bag of words (BoW) created with term frequency-inverse document frequency (TF-IDF), word embeddings using W2Vec and GloVe embeddings, logistic regression, support vector machine (SVM), and Naïve Bayes. They concluded that word embeddings have the highest performance for gender recognition. Ikae and

Savoy [10] compared different machine learning methods for gender detection using tweets including logistic regression, decision tree, k-nearest neighbors (KNN), SVM, Naïve Bayes, neural networks and random forest on seven different datasets. They concluded that neural networks and random forest perform best among the different approaches. Authors in [12] used n-grams as well as unigrams to tokenize sentences. They applied five different machine learning algorithms, Naive Bayes, sequential minimal optimization (SMO), logistic regression, random forest and j48 on text for gender recognition and found that a combination of 1- to 4-grams with SMO produces the best accuracy.

The studies mentioned above, have only used text for gender recognition and have not considered image data. Authors in [49] were the first to use profile images for gender detection. They stacked different approaches, namely, Microsoft discussion graph tool (DGT) using the username of the users, Face++ using their profile images, and SVMLight using their tweets. However, they combined pre-existing methods and did not train or fine-tune any model. In [44], VGG, a well-known image recognition model based on CNN has been fine-tuned for gender detection of Twitter users. In [50], text and image have been used for predicting the gender of Twitter users. In the image classification method, a CNN is trained for gender recognition. The text classification method includes applying TF-IDF to the hashtags and using latent Dirichlet allocation (LDA) to find the topics that the user is interested in. The results show that the combined method has higher accuracy.

Some studies have focused on image classification techniques for gender recognition. For example, authors in [51] propose a method for gender detection using images. First, they use CNN for feature extraction. Next, they apply a self-joint attention model for feature fusion. Finally, they use two fully connected neural network layers with ReLu and SoftMax activation functions and one average pooling layer to predict the gender. In [52], a method using gated residual attention networks has been proposed for gender recognition using images and tested on five different datasets. In [53], different CNNs are trained for gender recognition using different methods such as KNN, decision tree, SVM and SoftMax for feature extraction. The results of the CNN methods are combined by majority voting to increase the accuracy. Authors in [11] used posts, comments and replies on Facebook for gender recognition. They compared BERT with different machine learning and deep learning algorithms such as Naïve Bayes, Naïve Bayes Multinomial, SVM, decision tree, random forest, KNN, RNN and CNN. The results show that BERT has the highest performance among the different methods.

Some studies have combined both text and image classification models and employed transformers for gender recognition. Authors in [54] have designed a model for gender identification of Twitter users that combines three models, a multi-classifier for basic features (e.g. name, description), a multi-classifier for advanced features (i.e. k top words of tweets) and a ResNet-18 classifier for profile images of users. Among all the different methods (i.e. decision tree, SVM, AdaBoost, Gradient Boosting and Random Forest) that have been used for the multi-classifiers and for combining the models, Gradient Boosting has the highest accuracy. In [55], a multimodal approach using both text and image is proposed for the gender detection of Twitter users. The text classification part uses BERTBASE and the image classification part uses EfficientNet, a CNN-based approach for image recognition. The two methods are then combined to gain a higher accuracy. In [13], the gender of Twitter users has been predicted using their names, descriptions, tweets, and profile colors. SVM, BERT and BLSTM have been applied to user descriptions and BERT has performed better compared to SVM and BLSTM. Next, the different approaches are combined to improve the accuracy.

Some methods mentioned above have used transformers for text classification for author profiling, and have found that transformers have higher performance compared to other methods. However, transformer-based text-classification models have not been enhanced with transformer-based image recognition models for demographic information extraction. In this paper, we use transformer models to improve the performance and accuracy of text classification by combining it with image classification for gender recognition of Twitter users.

3. Materials and methods

Two datasets were used to conduct this study. The first dataset which was released in 2016 and is freely available on Kaggle includes the link to the profile image and one random tweet of 20050 different Twitter users [45]. The dataset has four different labels for the users, female, male, brand and unknown. The second dataset is PAN-18 which was released in 2018 and includes 100 tweets and 10 images posted by Arabic, English and Spanish speaking Twitter users. In this work, powerful models based on transformers were fine-tuned, tested, combined and compared on both of the datasets. In the following, the datasets, methods and models are explained.

3.1. The Kaggle dataset

The Kaggle dataset which could be downloaded from [45] includes link to profile image and a random tweet of 20050 different Twitter users. However, one single tweet does not carry much information and is not enough for training a strong gender classification model. Moreover, most of the profile image links did not work. Therefore, similar to other works [42,56], we gathered more tweets and the updated profile image link of the Twitter users of the Kaggle dataset. The tweet IDs of the dataset that we gathered is available at [48]. First, all the users with the unknown label were removed. Then, using the Twitter API Academic Researcher account and through the usernames of the users provided by the Kaggle dataset, user IDs and subsequently updated profile image links and approximately 100 different tweets posted by the user were gathered. The tweet IDs that were retrieved for each user are available online [48]. In compliance with Twitter's privacy policy, only the Tweet IDs and user IDs could be publicly released [57]. To obtain the text and other metadata, e.g. create date and location, the Tweet IDs need to be hydrated [58]. Tweets were cleaned, hyperlinks and mentions were removed and punctuations were fixed. Emojis were preserved since they carry valuable information that machine learning models could significantly benefit from. After balancing the dataset 2943 records of each class, i.e. female, male and brand, were acquired.

3.2. The PAN-18 dataset

In the PAN-18 dataset, which can be downloaded from [47] after permission is granted, 100 tweets and 10 images posted by 2500, 4900 and 5200 Arabic, English and Spanish speaking Twitter users have been gathered, respectively. Among the users 1000, 1900 and 2200 belong to the Arabic, English and Spanish testing and the rest to the training datasets, respectively. All the users have been labelled based on their gender, i.e. female and male. Half of the users in the training and testing datasets are female and the other half are male and the datasets are completely balanced. The tweets included emojis and were already cleaned. We used only the English datasets to train and test our models. Each of the 10 images of a user carries some information that could help the model separate the two genders. We found that by concatenating several images and feeding them into the base model for fine-tuning, the accuracy will significantly increase. The reason is that an image created from several images carries more information about the user and can help classify the gender with higher confidence. Since nine images can be concatenated to create a square image, nine of the ten images of a user were selected for concatenation. This was repeated ten times for a user, each time a different image was left out. The final image was resized to 224 × 224 pixels to be compatible with the transformer models. In order for our work to be reproducible, we have provided the code ("concatenate_images.py" in [48]) to generate the exact image combination that was used for training the models. We found that the accuracy of the model fine-tuned using the concatenated images is up to 16.92% higher compared to the model fine-tuned using the single original images.

3.3. Fine-tuning the image classification models

Deep learning models such as transformers are advantageous to other machine learning models only when a large dataset is fed to them. Oftentimes, labelled data is not available or it is very limited. In cases where a great amount of data is not accessible, fine-tuning a pre-trained deep learning model can help find the desired accuracy. To this end, we have fine-tuned ViT-Base, LeViT and Swin transformer for gender recognition of users based on their Twitter profile images (the Kaggle dataset) and based on ten different images that they have posted on Twitter (the PAN-18 dataset).

We split the Kaggle dataset into balanced train, validation, and test datasets with precisely, 7332, 498 and 999 users, respectively. Three models, namely, ViT, LeViT and Swin transformer, were fine-tuned to classify the images into three classes, female, male and brand.

For the PAN-18 dataset, 100 users of the train dataset were pulled out and allocated to the validation dataset. For each user ten concatenated images were created. All the ten images created for all the users from the training dataset were used for fine tuning the same three models (ViT, LeViT and Swin transformer). The accuracy of the model fine-tuned using the concatenated images was up to 16.92% higher compared to the model fine-tuned using the original images. Next, for each user the results of the ten concatenated images were combined using two fully connected neural network layers (Figure 1).

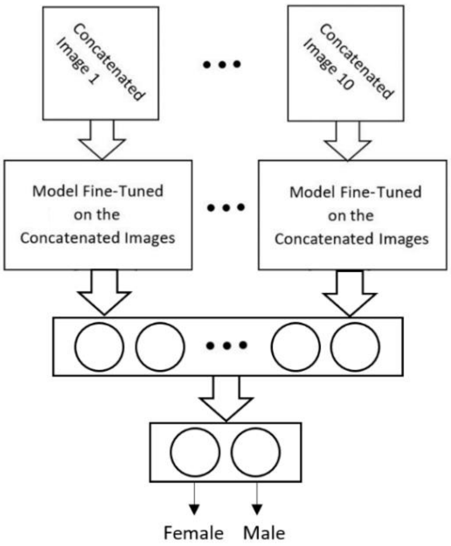

Concatenated Image 1
Concatenated Image 10
Model Fine-Tuned on the Concatenated Images
Model Fine-Tuned on the Concatenated Images
Female
Male

Figure 1. The image classification model for gender recognition of the PAN-18 dataset.

Before training each model, the cross-validation datasets were used for hyperparameter optimization using the WandB (weights and biases) library. We found that fine-tuning deep-learning models was not sensitive to the hyperparameters. The reason is that they have already been trained on a large dataset and require only a few more epochs to be fine-tuned. However, the fully connected neural network which combines the results of the ten different concatenated images was highly sensitive to the hyperparameters, since it was being trained from scratch. Most importantly, it was sensitive to the optimizer and performed well with Adam or AdamW optimizers, but very poor with the stochastic gradient descent (SGD) optimizer. Also, we found that smaller learning rates (≥ 0.001) work better when training the stacked neural networks. Table 1 shows the best hyperparameters used for training the stacked layers which combined the ten different concatenated images created for the PAN-18 dataset.

Table 1. The optimized hyperparameters of the stacked neural network trained for combining the image classification models of the PAN-18 dataset.

Model	Batch size	Dropout	Hidden layer size	Optimizer	Learning rate
ViT	16	0.5	8	AdamW	0.001
LeViT	16	0.2	8	AdamW	0.0001
Swin Transformer	16	0.2	5	AdamW	0.001

3.4. Fine-tuning the text-classification models

Some Twitter users may not have a suitable image for detecting their gender. However, we are able to retrieve the tweets of most Twitter users. Therefore, training a text classification model for gender recognition could help extract the gender of more users and increase the performance of the model. In both of our Kaggle and PAN-18 datasets, one hundred tweets are available for each user and are used to fine-tune the three transformer-based models, namely, BERTBASE, RoBERTa and ELECTRA for gender recognition.

We found that longer tweets result in a higher accuracy. Therefore, concatenating several tweets and using them for training the models significantly increase the accuracy. Since the number of tokens fed into BERT$_{\text{BASE}}$ cannot exceed 512, ten number of tweets could be concatenated at maximum. Thus, for each user we found ten concatenated tweets, and used them to fine-tune the models. This increased the accuracy of the model by 28.8 and 27.9% for the Kaggle and PAN-18 datasets, respectively. The model had three outputs, female, male and brand for the Kaggle dataset and two outputs, female and male for the PAN-18 dataset. The output of the model for each of the concatenated tweets of a user were combined using two fully connected layers. Figure 2 shows the text-classification model for the (A) Kaggle and (B) PAN-18 datasets. Similar to image-classification models, fine-tuning on text-classification models was not sensitive to hyperparameters. However, the stacked layer was highly sensitive to the hyperparameters, especially the optimizer and performed poorly with the SGD optimizer. Moreover, lower learning rates provided a higher accuracy. Table 2 shows the hyperparameters optimized using WandB library for the stacked fully connected network of the two datasets.

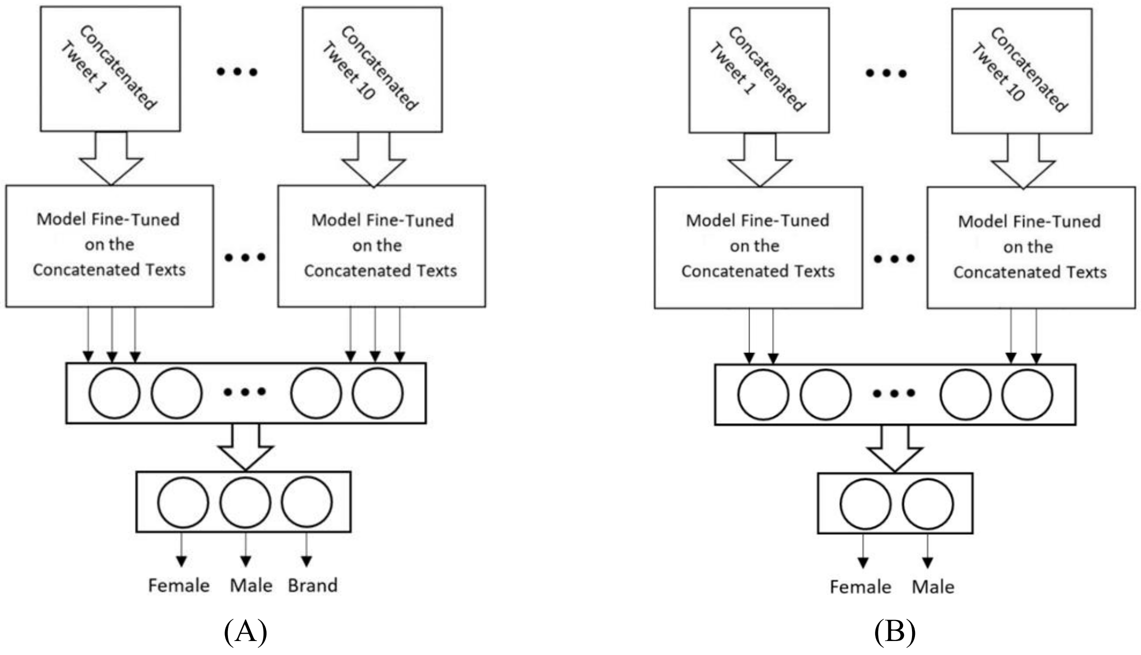

Concatenated Tweet 1
Concatenated Tweet 10
Model Fine-Tuned on the Concatenated Texts
Model Fine-Tuned on the Concatenated Texts
Female
Male
Brand
(A)
Concatenated Tweet 1
Concatenated Tweet 10
Model Fine-Tuned on the Concatenated Texts
Model Fine-Tuned on the Concatenated Texts
Female
Male
(B)

Figure 2. The text-classification model for gender recognition of the (A) Kaggle and (B) PAN-18 datasets.

Table 2. The optimized hyperparameters of the stacked neural network trained for combining the text classification models of the Kaggle and PAN-18 datasets.

	Model	Batch size	Dropout	Hidden layer size	Optimizer	Learning rate
The Kaggle dataset	BERT	16	0.1	10	Adam	0.001
	RoBERTa	16	0.1	5	Adam	0.001
	ELECTRA	16	0.2	10	Adam	0.001
The PAN-18 dataset	BERT	32	0.2	10	Adam	0.001
	RoBERTa	32	0.1	5	Adam	0.001
	ELECTRA	32	0.2	10	Adam	0.001

3.5. Combining text and image classification

For each of the Kaggle and PAN-18 datasets, the image and text classification models were combined using a neural network of two stacked layers. A SoftMax layer was placed at the top of the model to get the final outputs. Figures 3 shows the complete model for (A) the Kaggle and (B) the PAN-18 datasets. Since the model was built using three image-classification models and three text-classification models, nine different combinations were possible. Table 3 shows the optimized hyperparameters of the final stacked neural network for the nine different combinations. Our code is available at [48].

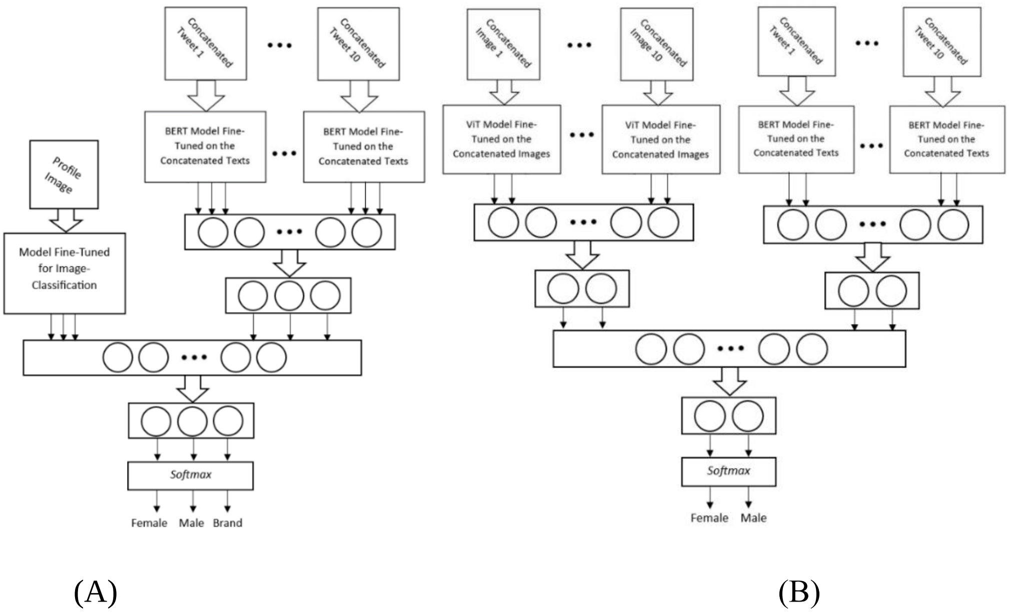

Profile Image
Model Fine-Tuned for Image-Classification
Concatenated Tweet 1
Concatenated Tweet 10
BERT Model Fine-Tuned on the Concatenated Texts
BERT Model Fine-Tuned on the Concatenated Texts
Softmax
Female
Male
Brand
(A)
Concatenated Image 1
Concatenated Image 10
ViT Model Fine-Tuned on the Concatenated Images
ViT Model Fine-Tuned on the Concatenated Images
Concatenated Tweet 1
Concatenated Tweet 10
BERT Model Fine-Tuned on the Concatenated Texts
BERT Model Fine-Tuned on the Concatenated Texts
Softmax
Female
Male
(B)

Figure 3. The complete model for (A) the Kaggle and (B) the PAN-18 datasets.

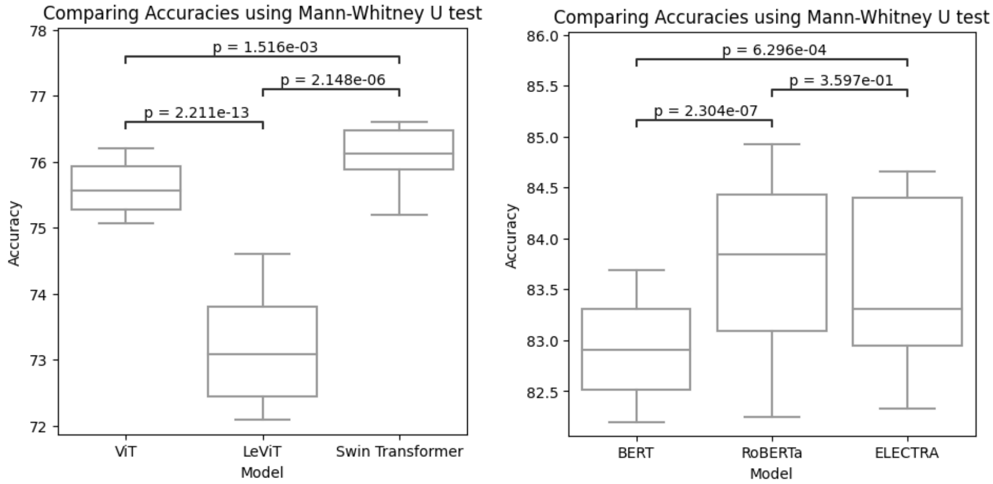
Table 3. The optimized hyperparameters for the nine different combinations of image- and text-classification for the Kaggle and PAN-18 datasets.

	Vision model	NLP Model	Batch size	Dropout	Hidden layer size	Optimizer	Learning rate
The Kaggle dataset	ViT	BERT	16	0.2	5	AdamW	10^{-5}
		RoBERTa	16	0.2	5	AdamW	10^{-5}
		ELECTRA	16	0.2	5	AdamW	10^{-5}
	LeViT	BERT	16	0.2	5	AdamW	10^{-5}
		RoBERTa	16	0.2	5	AdamW	10^{-5}
		ELECTRA	16	0.2	5	AdamW	10^{-5}
	Swin Transformer	BERT	16	0.5	5	AdamW	10^{-5}
		RoBERTa	16	0.5	5	AdamW	10^{-5}
		ELECTRA	16	0.5	5	AdamW	10^{-5}
The PAN-18 dataset	ViT	BERT	8	0.5	5	AdamW	10^{-5}
		RoBERTa	8	0.5	5	AdamW	10^{-5}
		ELECTRA	8	0.5	5	AdamW	10^{-5}
	LeViT	BERT	16	0.2	8	AdamW	10^{-5}
		RoBERTa	16	0.2	8	AdamW	10^{-5}
		ELECTRA	16	0.2	8	AdamW	10^{-5}
	Swin Transformer	BERT	16	0.5	5	AdamW	10^{-5}
		RoBERTa	16	0.5	5	AdamW	10^{-5}
		ELECTRA	16	0.5	5	AdamW	10^{-5}

4. Results

4.1. The Kaggle dataset

Each of the image-classification models were fine-tuned ten times, so that their statistical significance could be evaluated and compared. The different text-classification models were trained and built ten times, as well. Figure 4 compares the statistical significance of different (A) image- and (B) text-classification models through Mann-Whitney U test. In most cases, a p-value lower than 0.05 is considered significant in statistical analysis [59]. Table 4 compares the maximum accuracies of different models and their precision, recall and f1-scores. The p-value in Figure 4(A) indicates that the accuracy of the LeViT model is significantly lower than that of the ViT and Swin transformer models. This result suggests that transformer models that are enhanced with CNN have a lower accuracy compared to models that are solely built using transformers for our dataset. According to Table 4, Swin transformer provides a higher accuracy compared to the ViT and LeViT models on the Kaggle dataset. The same result is confirmed by Table 4 for the three image-classification models. The p-value in Figure 4(B) shows that the accuracy of BERT is significantly lower than RoBERTa and ELECTRA. Nonetheless, the accuracy of RoBERTa and ELECTRA are not significantly different from each other. However, Table 4 shows that the maximum accuracy of RoBERTa is higher than that of ELECTRA for the Kaggle dataset. Moreover, according to Table 4, the maximum accuracy of BERT is lower than RoBERTa and ELECTRA.



(A) (B)

Figure 4. The accuracy of the different (A) image- and (B) text-classification models compared using the Mann-Whitney U test.

The three different vision models were combined with the three different NLP models to acquire a higher accuracy through a multimodal approach. Table 5 compares the maximum accuracies and the precision, recall and f1-score of the nine different combinations with each other. Precision indicates the percentage of the correctly classified items detected for a particular class and recall indicates the percentage of the items from a particular class that were actually detected. High precision and recall for all the classes of the final combined model indicate that it can distinguish between all the classes pretty well. According to Table 5, the highest accuracy is obtained when the result of the Swin transformer model is combined with the NLP models and the best accuracy is acquired from the combination of Swin transformer and BERT. The accuracy of all the nine different combination models is higher than that of their image- and text-classification models. The accuracy of the Swin transformer-BERT multimodal is 11.73 and 5.26% higher than the accuracy of Swin transformer and RoBERTa models, respectively.

Table 4. The maximum accuracy and model evaluation parameters obtained for different models for the Kaggle dataset.

	Model	Class	Accuracy	Precision	Recall	F1-score
Image-classification models	ViT	Female	76.87	76.05	79.11	77.54
		Male		77.44	76.82	77.13
		Brand		75.93	75.34	75.63
	LeViT	Female	72.8	76.82	75.12	75.96
		Male		70.18	70.68	70.43
		Brand		75.69	74.22	74.94
	Swin Transformer	Female	**78.86**	79.92	75.32	77.55
		Male		81.61	74.42	77.85
		Brand		74.12	79.09	76.52
Text-classification models	BERT	Female	83.69	82.11	83.25	82.68
		Male		82.02	83.17	82.59
		Brand		85.54	81.93	83.7
	RoBERTa	Female	**84.92**	83.11	85	84.04
		Male		83.09	85.21	84.14
		Brand		85.91	83.11	84.49
	ELECTRA	Female	84.66	82.64	85.78	84.18
		Male		82.81	85.16	83.97
		Brand		84.81	82.04	86.35

Table 5. The maximum accuracy and model evaluation parameters of the multimodal techniques for the Kaggle dataset.

Vision	NLP	Class	Accuracy	Precision	Recall	F1-score
ViT	BERT	Female	82.84	79.80	83.74	81.72
		Male		80.09	83.22	81.63
		Brand		86.13	80.13	83.02
	RoBERTa	Female	83.42	81.12	85.67	83.33
		Male		80.79	85.21	82.94
		Brand		86.49	81.52	83.93
	ELECTRA	Female	83.11	80.02	86.04	82.92
		Male		80.13	85.86	82.9
		Brand		86.17	80.42	83.2

Continue to next page

Vision	NLP	Class	Accuracy	Precision	Recall	F1-score
LeViT	BERT	Female	81.47	77.23	83.15	80.08
		Male		76.49	83.41	79.8
		Brand		84.33	78.12	81.11
	RoBERTa	Female	81.79	78.71	84.39	81.45
		Male		78.48	84.84	81.53
		Brand		85.12	78.97	81.93
	ELECTRA	Female	81.56	76.81	85.93	81.11
		Male		77.28	86.10	81.45
		Brand		86.33	78.51	82.23
Swin Transformer	BERT	Female	**88.11**	85.81	90.42	88.05
		Male		86.14	89.2	87.64
		Brand		91.39	86.12	88.68
	RoBERTa	Female	85.32	82.23	88.11	85.07
		Male		81.91	87.87	84.78
		Brand		89.19	82.18	85.54
	ELECTRA	Female	85.74	82.48	88.32	85.3
		Male		82.14	88.48	85.19
		Brand		88.94	82.11	85.89

4.2. The PAN-18 dataset

Similar to the Kaggle dataset, each image- and text-classification model for the PAN-18 was built and tested ten different times. Figure 5 evaluates the statistical significance of the (A) image- and (B) text-classification models using the Mann-Whitney U test. Moreover, Table 6 compares the maximum accuracies of different vision and NLP models. Figure 5(A) shows that the accuracy of the Swin transformer model is significantly higher than the other two models and the accuracy of ViT is significantly higher than LeViT model. Additionally, according to Table 6, the maximum accuracy observed for Swin transformer is higher than ViT and LeViT and the maximum accuracy observed for ViT is higher than LeViT. Figure 5(B) shows that RoBERTa has a significantly higher accuracy compared to ELECTRA and BERT, but BERT and ELECTRA are not significantly different for the PAN-18 dataset. However, according to Table 6, the maximum accuracy of RoBERTa is higher than BERT and ELECTRA and the maximum accuracy of ELECTRA is higher than BERT.

Maximum accuracies, and their precision, recall, and f1-score of the nine different multimodal methods for the PAN-18 dataset are compared in Table 7. Table 7 shows that the final model has a high value for the precision, recall and f1-score for the two female and male classes. This means that the model performs well for both of the classes and is capable of distinguishing them from each other very well. In addition, the maximum accuracy of the models combined with Swin transformer is dominantly higher than that of other models. Although RoBERTa significantly had a higher accuracy compared to other NLP models (Figure 5(B)), the best accuracy was obtained when the Swin transformer and BERT were combined. The maximum accuracy of the combination of Swin transformer-BERT is 8.55 and 9.8% higher than that of the Swin transformer and BERT models, respectively.

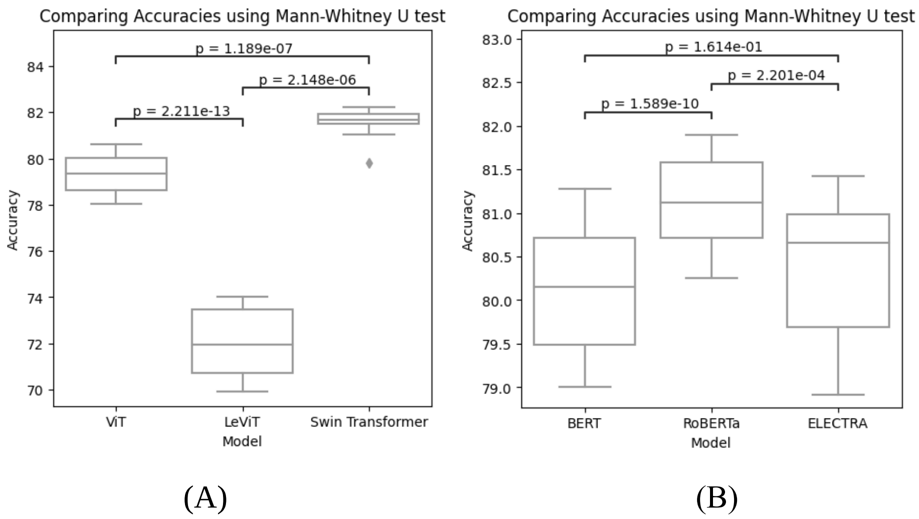


Figure 5. Comparing different (A) image- and (B) text-classification models using Mann-Whitney U test.

Table 6. The maximum accuracy and model evaluation parameters obtained for different models for the PAN-18 dataset.

	Model	Class	Accuracy	Precision	Recall	F1-score
Image-classification models	ViT	Female	80.82	81.46	78.75	80.8
		Male		79.11	82.27	80.65
	LeViT	Female	74.22	73.76	75.53	74.63
		Male		74.89	73.11	73.99
	Swin Transformer	Female	**82.21**	83.70	80	81.81
		Male		80.84	84.42	82.59
Text-classification models	BERT	Female	81.27	79.98	83.71	81.80
		Male		83.08	79.05	81.02
	RoBERTa	Female	**81.89**	80.54	84.11	82.29
		Male		83.37	79.68	81.48
	ELECTRA	Female	81.42	79.24	85.16	82.09
		Male		83.96	77.68	80.7

Table 7. The maximum accuracy and model evaluation parameters of the multimodal techniques for the PAN-18 dataset.

Vision	NLP	Class	Accuracy	Precision	Recall	F1-score
ViT	BERT	Female	86.79	87.01	86.92	86.97
		Male		85.63	87.12	86.37
	RoBERTa	Female	85.39	83.82	88.06	85.89
		Male		87.34	83.24	85.24
	ELECTRA	Female	85.48	83.03	87.74	85.32
		Male		88.31	82.96	85.55

Continue to next page

Vision	NLP	Class	Accuracy	Precision	Recall	F1-score
LeViT	BERT	Female	76.87	78.41	74.21	76.25
		Male		74.88	79.09	76.92
	RoBERTa	Female	79.42	82.41	78.12	80.21
		Male		77.33	81.17	79.2
	ELECTRA	Female	78.91	80.89	77.64	79.23
		Male		76.43	79.14	77.76
Swin Transformer	BERT	Female	**89.24**	91.27	88.12	89.66
		Male		87.49	90.95	89.18
	RoBERTa	Female	88.36	90.13	86.97	88.52
		Male		86.73	89.86	88.26
	ELECTRA	Female	88.22	89.93	86.92	88.4
		Male		87.01	89.14	88.06

4.3. Comparing with other models

Table 8 compares the RoBERTa text-based model with the work done in [56] for the Kaggle dataset, and the RoBERTa text-based model and the Swin transformer-BERT multimodal with the works done in [15,43,60], which had the first, second, and third ranks in the PAN author profiling competition of 2018, for the PAN-18 dataset.

Authors in [56] have used the Kaggle dataset to build a text-based gender recognition model. After retrieving additional tweets for the female and male users, the tweets were cleaned and vectorized using a number of methods, namely, BOW, TF-IDF, Word2vec, GLObal VEctor for word representation (GLOVE) and BERT tokenization. Different machine learning algorithms were used to build a gender recognition model. Best results were obtained using GLOVE and random forest (RF) and GLOVE and SVM. We applied the GLOVE-RF and GLOVE-SVM models on our dataset and compared it with the RoBERTa text-classification model for only male and female classes (Table 8). Authors in [15] have used bidirectional GRU for text classification and CNN based on VGG16 for image classification parts. Then, the image and text classification parts are combined using fusion component which includes direct multiplication of text and image feature components. In [60], authors classified Twitter users using only text. They applied TF-IDF and singular value decomposition (SVD) on the tweets to extract the semantics. Then they applied latent semantic analysis (LSA) to extract the semantic topics and fed them into an SVM with linear kernel for gender classification. Authors in [43] proposed an approach for gender identification of PAN-18 dataset using text and image. They applied TF-IDF and then SVD to extract the semantics and use them for gender classification using linear-SVM. For image classification, they stacked three different classification layers. The first layer, low classifier, consisted of four different classifiers, object recognition, facial recognition, color histogram and local binary patterns. They all used linear-SVC except for color histogram that used multinomial naïve bayes (NB). The second classifier layer, meta-classifier used linear-SVC to combined the results of the four classifiers of the previous layer. The third classifier layer, aggregation classifier, combined the meta-classifier results of the ten different images of a user using MultinomialNB. Finally, they combined their text and image classifiers using linear-SVC. Table 2 shows that our multimodal method is superior to all the above models in terms of accuracy.

Table 8. Comparing the Swin Transformer-BERT model with other methods in terms of accuracy.

		Text-based	Image-based	Overall
The Kaggle dataset	Text-based with RF [56]	71.22%	-	-
	Text-based with SVM [56]	69.14%	-	-
	Our Model (RoBERTa)	84.09%	-	-
The PAN-18 dataset	Multimodal [44]	79.68%	81.63%	85.84%
	Text-based [60]	82.21%	-	-
	Multimodal [43]	80.74%	69.63%	81.32%
	Our Model (RoBERTa)	**81.89%**	-	-
	Our Model (Swin Transformer-BERT)	81.27%	**82.21%**	**89.24%**

5. Discussion

Previously, some works have used profile images of social media users and some other have used their image contents posted on social media for gender recognition. In this work, we have implemented a transformer-based model for both of the methods. We extended a publicly available dataset for gender recognition with profile images and used the PAN-18 dataset for gender recognition with image content posted on social media. Our results show that using the image content posted by users on social media a higher accuracy is obtained.

To further improve the accuracy of our model, the image-classification model was combined with a text-classification. Different transformer-based image classification models, namely, ViT, LeViT and Swin transformer and text classification models, i.e. BERT, RoBERTa and ELECTRA were explored. Swin transformer dominantly performed better than other vision models for both of the datasets. In contrast, LeViT had a lower accuracy compared to other models on our dataset. This shows that models built solely on transformers have a higher accuracy compared to models enhanced with CNN for our datasets. RoBERTa had a significantly higher accuracy compared to BERT for both of the datasets. However, BERT performed better when combined with Swin transformer. BERT and Swin transformer complemented each other very well and provided the best accuracy of 88.11% and 89.24% for the Kaggle and PAN-18 datasets.

One limitation to our work was lack of suitable dataset. To remove this barrier, we have completed the publicly available dataset on Kaggle and provided approximately 100 tweet IDs for female, male and brand classes that future studies could build up on.

6. Conclusions and future work

Demographics of social media users are beneficial for research and applications in health, socio-economic inequalities and gender vulnerability. However, such information is not usually and freely available. During periods of upheaval, women are usually at greater risk from the adverse effects and potential losses incurred by these external stressors. They are also the slowest to recover from such emergencies. Integrating governance at widening levels and mitigating the limited economic options of women, are two examples of systemic challenges which require attention for human futurity. However, in many cases, even the data required to document and understand these challenges is not available. This paper addresses these systemic imperatives by providing a framework that can help us to identify the elements of promising emergent governance frameworks to address local and global-scale socio-economic challenges that disproportionately impact women.

In this work we have designed a model based on transformers to detect the gender of Twitter users using both text and image. We have implemented and compared our multimodal method using several transformer models and found that the combination of Swin transformer and BERT complement each other better provides the best accuracy for our datasets.

Future studies could build on our work by using other user information such as descriptions, media posts, comments and likes. Moreover, recognizing other user demographics such as age and ethnicity using transformers could be further investigated. In addition, heuristic methods for identifying user demographics when images are blurry, have low quality, are partially viewed or when people are wearing masks or sunglasses can be studied for higher accuracy and better performance.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgment

This research is funded by Canada's International Development Research Centre (IDRC) and Swedish International Development Cooperation Agency (SIDA) (Grant No. 109559-001).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. J. Gao, P. Zheng, Y. Jia, H. Chen, Y. Mao, S. Chen, et al., Mental health problems and social media exposure during COVID-19 outbreak, *PLOS ONE*, **15** (2020). https://doi.org/10.1371/journal.pone.0231924
2. M. J. Aramburu, R. Berlanga, I. Lanza, Social media multidimensional analysis for intelligent health surveillance, *Int. J. Env. Res. Public Health*, **17** (2020), 2289. https://doi.org/10.3390/ijerph17072289
3. J. B. Whiting, J. C. Pickens, A. L. Sagers, M. PettyJohn, B. Davies, Trauma, social media, and #WhyIDidntReport: An analysis of twitter posts about reluctance to report sexual assault, *J. Marital. Fam. Ther.*, **47** (2021), 749–766. https://doi.org/10.1111/jmft.12470
4. T. Simon, A. Goldberg, L. Aharonson-Daniel, D. Leykin, B. Adini, Twitter in the cross fire–the use of social media in the Westgate Mall terror attack in kenya, PLOS ONE, **9** (2014). https://doi.org/10.1371/journal.pone.0104136
5. G. Coppersmith, R. Leary, A. Fine, Natural language processing of social media as screening for suicide risk, *Biomed. Inform. Insights*, **10** (2018). https://doi.org/10.1177/1178222618792860
6. S. S. Hill, F. J. Dore, T. E. Steven, R. J. McLoughlin, A. S. Crawford, P. R. Sturrock, et al., Twitter use among departments of surgery with general surgery residency programs, *J. Surg. Educ.*, **78** (2021), 35–42. https://doi.org/10.1016/j.jsurg.2020.06.008

7. K. R. Blake, B. Bastian, T. F. Denson, R. C. Brooks, Income inequality not gender inequality positively covaries with female sexualization on social media, *Proc. Natl. Acad. Sci. U. S. A.*, **115** (2018), 8722–8727. https://doi.org/10.1073/pnas.1717959115
8. S. Ahemd, D. Madrid-Morales, Is it still a man's world? Social media news and gender inequality in online political engagement, *Inform. Commun. Soc.*, **24** (2020), 381–399. https://doi.org/10.1080/1369118X.2020.1851387
9. P. Vashisth, K. Meehan, Gender classification using Twitter text data, in *2020 31st Irish Signals and Systems Conference (ISSC)*, (2020), 1–6. https://doi.org/10.1109/ISSC49989.2020.9180161
10. C. Ikae, J. Savoy, Gender identification on Twitter, *J. Assoc. Inform. Sci. Tech.*, **73** (2021), 58–69. https://doi.org/10.1002/asi.24541
11. Ö. Çoban, A. İnan, S. A. Özel, Facebook tells me your gender: An exploratory study of gender prediction for Turkish Facebook users, *ACM Trans. Asian Low-Reso.*, **20** (2021), 1–38. https://doi.org/10.1145/3448253
12. I. Ameer, G. Sidorov, R. M. A. Nawab, Author profiling for age and gender using combination of features of various types, *J. Intell. Fuzzy Syst.*, **36** (2019), 4833–4843. https://doi.org/10.3233/JIFS-179031
13. Y. C. Yang, M. A. Al-Garadi, J. S. Love, J. Perrone, A. Sarker, Automatic gender detection in Twitter profiles for health-related cohort studies, *JAMIA Open*, **4** (2021). https://doi.org/10.1093/jamiaopen/ooab042
14. C. Suman, A. Naman, S. Saha, P. Bhattacharyya, A multimodal author profiling system for tweets, *IEEE Trans. Comput. Social Syst.*, **8** (2021), 1407–1416. https://doi.org/10.1109/TCSS.2021.3082942
15. T. Takahashi, T. Tahara, K. Nagatani, Y. Miura, T. Taniguchi, T. Ohkuma, Text and image synergy with feature cross technique for gender identification, in *Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, (2018), 1–22.
16. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, (2017), 6000–6010.
17. Y. Kim, C. Denton, L. Hoang, A. M. Rush, Structured attention networks, preprint, arXiv:1702.00887.
18. A. Parikh, O. Täckström, D. Das, J. Uszkoreit, A decomposable attention model for natural language inference, preprint, arXiv:1606.01933.
19. A. Galassi, M. Lippi, P. Torroni, Attention in natural language processing, *IEEE Trans. Neur. Net. Lear. Syst.*, **32** (2021), 4291–4308. https://doi.org/10.1109/TNNLS.2020.3019893
20. P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, J. Shlens, Stand-Alone Self-Attention in Vision Models, *Adv. Neur. Inform. Process. Syst.*, **32** (2019).
21. J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, **1** (2019), 4171–4186. https://doi.org/10.18653/v1/N19-1423
22. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16 × 16 words: Transformers for image recognition at scale, preprint, arXiv:2010.11929.
23. OpenAI, Improving language understanding with unsupervised learning, 2018, [cited 19 June 2023]. Available from: https://openai.com/research/language-unsupervised
24. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, et al., RoBERTa: A robustly optimized BERT pretraining approach, preprint, arXiv: cs.CL/1907.11692.

25. Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. V. Le, XLNet: Generalized autoregressive pretraining for language understanding, preprint, arXiv:cs.CL/1906.08237.
26. K. Clark, M. T. Luong, Q. V. Le, C. D. Manning, ELECTRA: Pre-training text encoders as discriminators rather than generators, preprint, arXiv:cs.CL/2003.10555v1.
27. M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, et al., Generative pretraining from pixels, in *Proceedings of the 37th International Conference on Machine Learning*, **119** (2020), 119–1691.
28. Google Research, Google Brain Team, 2011, [cited 24 May 2022]. Available from: https://research.google/teams/brain/
29. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90
30. M. Tan, Q. V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, preprint, arXiv:cs.LG/1905.11946.
31. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jegou, Training data-efficient image transformers & distillation through attention, preprint, arXiv:cs.CV/2012.12877.
32. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, et al., Swin transformer: Hierarchical vision transformer using shifted windows, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, ICCV, (2021), 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986
33. K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, W. Wu, Incorporating convolution designs into visual transformers, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 559–568. https://doi.org/10.1109/ICCV48922.2021.00062
34. Y. Li, K. Zhang, J. Cao, R. Timofte, L. V. Gool, LocalViT: Vringing locality to vision transformers, preprint, arXiv:cs.CV/2104.05707.
35. A. Srinivas, T. Y. Lin, N. Parmar, J. Shlens, P. Abbeel, A. Vaswani, Bottleneck transformers for visual recognition, preprint, arXiv:cs.CV/2101.11605.
36. H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, et al., CvT: Introducing convolutions to vision transformers, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021). https://doi.org/10.1109/ICCV48922.2021.00009
37. B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jegou, et al., LeViT: a Vosopm transformer in ConvNet’s clothing for faster inference, preprint, arXiv:cs.CV/2104.01136.
38. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, *Computer Vision – ECCV 2020*, Springer, Cham, (2020). https://doi.org/10.1007/978-3-030-58452-8_13
39. Y. Fang, B. Liao, X. Wang, J. Fang, J. Qi, R. Wu, et al., You only look at one sequence: Rethinking transformer in vision through object detection, preprint, arXiv:cs.CV/2106.00666.
40. H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, L. C. Chen, Axial-DeepLab: Stand alone axial-attention for panoptic segmentation, *Computer Vision – ECCV 2020*, Springer, Cham, (2020). https://doi.org/10.1007/978-3-030-58548-8_7
41. Y. Jiang, S. Chang, Z. Wang, TransGAN: Two pure transformers can make one strong GAN, and that can scale up, preprint, arXiv:cs.CV/2102.07074.
42. L. Li, Z. Song, X. Zhang, E. A. Fox, A hybrid model for role-related user classification on Twitter, preprint, arXiv:cs.SI/1811.10202.

43. G. Ciccone, A. Sultan, L. Laporte, E. Egyed-Asigmond, A. Alhamzeh, M. Granitzer, Stacked gender prediction from tweet texts and images notebook for pan at CLEF 2018, in *CLEF 2018-Conference and Labs of the Evaluation*, (2018).
44. M. A. Alvarez-Carmona, L. Pellegrin, M. Montes-y-Gómez, F. Sánchez-Vega, H. J. Escalante, A. P. López-Monroy, et al., A visual approach for age and gender identification on Twitter, *J. Intell. Fuzzy Syst.*, **34** (2018), 3133–3145. https://doi.org/10.3233/JIFS-169497
45. Twitter User Gender Classification, DATA, 2016, [cited 21 June 2023]. Available from: https://www.kaggle.com/datasets/crowdflower/twitter-user-gender-classification?resource=download
46. F. Rangel, P. Rosso, M. M-Y-Gómez, M. Potthast, B. Stein, Overview of the 6th author profiling task at pan 2018: Multimodal gender identification in Twitter, in *Working notes papers of the CLEF*, (2018).
47. PAN, DATA, 2018, [cited 21 June 2022]. Available from: https://pan.webis.de/data.html
48. Gender Recognition Using Transformers, 2023. Available from: https://github.com/Zahra1221/Gender-Recognition-using-Transformers
49. M. Sayyadiharikandeh, G. L. Ciampaglia, A. Flammini, Cross-domain gender detection in Twitter, in *Proceedings of the Workshop on Computational Approaches to Social Modeling*, (2016).
50. L. Geng, K. Zhang, X. Wei, X. Feng, Soft biometrics in online social networks: A case study on Twitter user gender recognition, in *2017 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, (2017), 1–8. https://doi.org/10.1109/WACVW.2017.8
51. X. Zhang, S. Javid, J. Dias, N. Werghi, Person gender classification on RGB-D data with self-joint attention, *IEEE Access*, **9** (2021), 166303–166313. https://doi.org/10.1109/ACCESS.2021.3135428
52. A. Garain, B. Ray, P. K. Singh, A. Ahmadian, N. Senu, R. Sarkar, GRA_NET: A deep learning model for classification of age and gender from facial images, *IEEE Access*, **9** (2021), 85672–85689. https://doi.org/10.1109/ACCESS.2021.3085971
53. J. Cheng, Y. Li, J. Wang, L. Yu, S. Wang, Exploiting effective facial patches for robust gender recognition, IEEE, *Tsinghua Sci. Technol.*, **24** (2019), 333–345. https://doi.org/10.26599/TST.2018.9010090
54. L. Li, Z. Song, X. Zhang, E. A. Fox, A hybrid model for role-related user classification on Twitter, preprint, arXiv:1811.10202.
55. C. Suman, A. Naman, S. Saha, P. Bhattacharyya, A multimodal author profiling system for tweets, *IEEE Trans. Comput. Social Syst.*, **8** (2021), 1407–1416. https://doi.org/10.1109/TCSS.2021.3082942
56. B. Onikoyi, N. Nnamoko, I. Korkontzelos, Gender prediction with descriptive textual data using a Machine Learning approach, *Natural Language Proces. J.*, **4** (2023). https://doi.org/10.1016/j.nlp.2023.100018
57. Twitter Developer Platform, Developer Agreement and Policy, 2023, Available from: https://developer.twitter.com/en/developer-terms/agreement-and-policy, (accessed 21 June 2023)
58. DocNow Hydrator, 2021, [cited 21 June 2023]. Available from: https://github.com/DocNow/hydrator
59. M. Jafari, N. Ansari-Pour, Why, When and How to Adjust Your P Values?, *Cell J.*, **20** (2019), 604–607. https://doi.org/10.22074/cellj.2019.5992
60. S. Daneshvar, D. Inkpen, Gender Identification in Twitter using N-grams and LSA, in *proceedings of the ninth international conference of the CLEF association (CLEF 2018)*, (2018).