Pew Research Center

FOR RELEASE MAY 23, 2019

Men Appear Twice as Often as Women in News Photos on Facebook

Photos that exclusively show men make up the majority of photos that show people; representational differences persist across topics

BY *Onyi Lam, Stefan Wojcik, Adam Hughes and Brian Broderick*

FOR MEDIA OR OTHER INQUIRIES:

Stefan Wojcik, Computational Social Scientist
Adam Hughes, Computational Social Scientist
Rachel Weisel, Communications Manager

202.419.4372

www.pewresearch.org

RECOMMENDED CITATION

Pew Research Center, May, 2019, "Men Appear Twice as Often as Women in News Photos on Facebook"

About Pew Research Center

Pew Research Center is a nonpartisan fact tank that informs the public about the issues, attitudes and trends shaping the world. It does not take policy positions. The Center conducts public opinion polling, demographic research, content analysis and other data-driven social science research. It studies U.S. politics and policy; journalism and media; internet, science and technology; religion and public life; Hispanic trends; global attitudes and trends; and U.S. social and demographic trends. All of the Center's reports are available at www.pewresearch.org. Pew Research Center is a subsidiary of The Pew Charitable Trusts, its primary funder.

Men Appear Twice as Often as Women in News Photos on Facebook

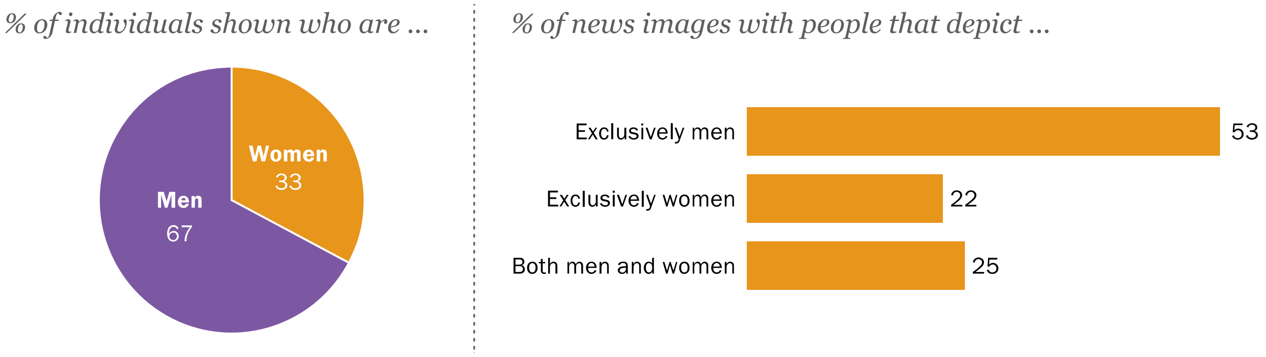
Photos that exclusively show men make up the majority of photos that show people; representational differences persist across topics

A new study of the images accompanying news stories posted publicly on Facebook by prominent American news media outlets finds that men appear twice as often as women do in news images, with a majority of photos showing exclusively men.

The Pew Research Center analysis used machine vision to examine the representation of women in news images from 17 national news outlets' Facebook posts from April 1 through June 30, 2018.[1] Facebook is a source of news for 43% of U.S. adults, while social media as a whole has outpaced print newspapers when it comes to where people often get their news.

Men appear more than women in news images on Facebook when it comes to both individuals and groups of people

Among images on Facebook from 17 national news organizations …


% of individuals shown who are …

% of news images with people that depict …

Source: Pew Research Center analysis of Facebook news images from 17 national news outlets created April 1–June 30, 2018.
"Men Appear Twice as Often as Women in News Photos on Facebook"

PEW RESEARCH CENTER

[1] News outlets use Facebook to post links (with photos) to their own articles. See Methodology for more information about how the outlets were selected and how their posts were collected. The time period covered by this study includes events such as the UK royal wedding and the World Cup.

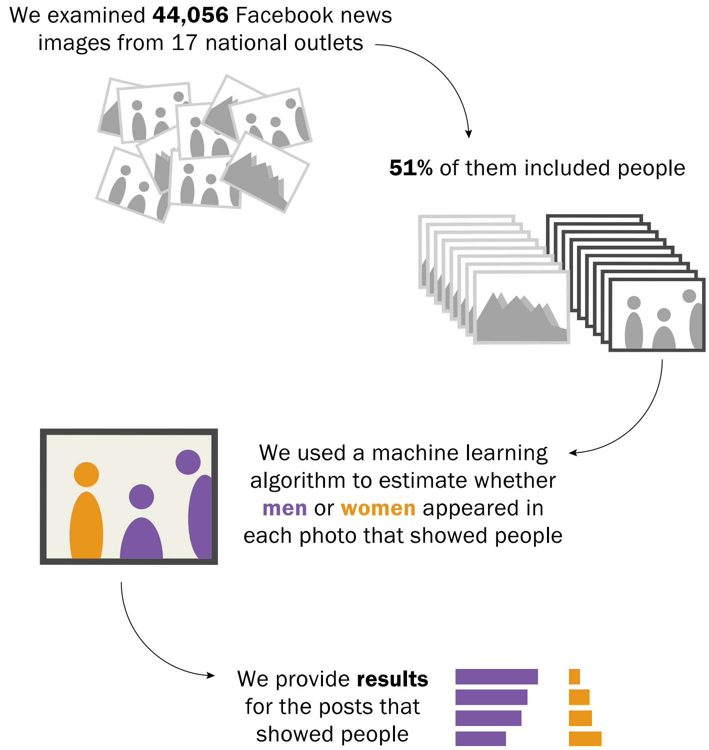
Why Pew Research Center studied news photos on Facebook

Researchers chose to study news images on Facebook because the site standardizes the presentation of news images and text across outlets. News posts that appear in social media feeds like Facebook feature large photographs and contain only a small amount of text and a link to a longer article. In contrast to other formats such as print media, the photograph in a Facebook post occupies more screen space than the accompanying text and is the main object that Facebook users see when they scroll through the news feed. Academic research based on data collected between July 2014 and January 2015 finds that Facebook users only clicked on about 7% of national news, politics and world affairs posts that they viewed in their news feeds. Previous research from the Center has examined how representations of men and women in Google Image Search results can sometimes be at odds with real world data. And academic researchers have leveraged similar tools to study the depiction of women in the news.

There are several ways to measure how often men and women appear in news photos. One way is to think about all the photos together as making up one big crowd of people and estimating what share are women versus men. Women made up 33% of all the 53,067 individuals identified in news post images, while men made up the other 67%.

Another way to look at the data is to examine the mix of people who appear in each image. Across the 22,342 posts with photos containing identifiable human faces, more than half of them (53%) exclusively showed men, while less than a quarter showed exclusively women (22%). The remaining images (25%) show at least one man and one woman.

How Pew Research Center examined gender representation in Facebook posts from news outlets



We examined **44,056** Facebook news images from 17 national outlets

51% of them included people

We used a machine learning algorithm to estimate whether **men** or **women** appeared in each photo that showed people

We provide **results** for the posts that showed people

“Men Appear Twice as Often as Women in News Photos on Facebook”

PEW RESEARCH CENTER

All 17 news outlets included in the study showed more men than women in news images on Facebook during the study period. The share of individuals who were identified as women by the model ranges from 25% to 46%, by outlet.

While these findings are striking, there is no perfect benchmark or "true ratio" for how often men and women *should* be portrayed in news images on social media. Yes, the U.S. population is divided nearly in half, male versus female. But, for example, all the representational coverage of professional football teams would return results overwhelmingly dominated by male faces. Coverage of the U.S. Senate – currently 25% female – might do the same. In addition, the analysis did not address whether the content of the news stories that accompanied the images was more focused on men or women.

The analysis also reveals other ways that men are more prominent in news images on Facebook. In photos that showed two or more people, men tend to outnumber women. And men's faces take up more space when shown, with the average male face being 10% larger than the average female face across all photos with people.

Selecting media outlets

The 17 media organizations included in the study were selected according to several criteria. These included: whether they conduct original reporting on general topics, whether they primarily covered national news rather than local news, whether the site was for a news organization based in the U.S., and whether their websites received at least 20 million unique visitors in the third quarter of 2018, according to data from Comscore. The study does not include media outlets that focus their coverage on one topic, such as business, politics, entertainment or sports. The study also excludes local media outlets. The full list of sites included in the study appears in the Methodology.

Across several different types of news content on Facebook, women appear in images at a consistently lower rate than men. In posts related to the economy, 9% of images that show people exclusively show women, compared with 69% of images which exclusively showed men. A total of 22% of individuals shown in stories about the economy were women, while 78% were men. For posts about entertainment, women made up 40% of depicted individuals and 27% of news photos exclusively showed women, compared with 42% for photos that showed exclusively men.

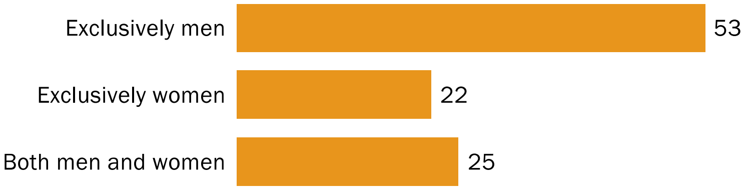
This work is part of the Center's exploration of the ways that new advances in machine vision allow researchers to use a computational model to analyze large quantities of photos. While the technology is not perfect – performance may not be as good as human judgment and could suffer from inconsistency across demographic groups (see Methodology for details) – it holds promise in analyzing a large quantity of data online, especially because researchers can measure how well these models perform directly. Overall, researchers used a model that achieved 95% accuracy when tested on a subset of the data it was trained with, and 87% accuracy when tested on a random sample of 998 individuals from the Facebook news images, using human judgments as a ground truth.

Men appear more than women in news photos on Facebook

A sizable proportion of the Facebook news images posted by the outlets examined in this study exclusively depict men. Across the news images that showed people, 22% showed exclusively women and 53% showed exclusively men.[2] Photos including both men and women were slightly more common (25%) than those that showed women alone.

Facebook news images more likely to show exclusively men than both men and women or exclusively women

% of news images with people from 17 national news organizations on Facebook that depict …



Source: Pew Research Center analysis of Facebook news images from 17 national news outlets created April 1-June 30, 2018.
"Men Appear Twice as Often as Women in News Photos on Facebook"

PEW RESEARCH CENTER

Overall, researchers identified 53,067 individuals across the 22,342 images from news posts showing people. Of those individuals, 35,367 were estimated to be men and 17,700 were estimated to be women. In other words, 33% of all people shown in news images on Facebook were women and 67% were men.

Over three-quarters of images showed one or more men (78%), while slightly fewer than half (47%) of photos showed at least one woman.

In photos that depicted multiple people, men outnumbered women. The median post with multiple people showed one woman and two men.

[2] The gender classification model makes predictions for men and women but does not include estimates for nonbinary individuals. More information about the classification model is available in the Methodology.

Measuring how often men and women appear across photos with multiple people

Calculating the percent of men or women who appear in news photos – among all depicted individuals – is a straightforward way of looking at the data. However, some photos show just one person, while others show many people. Calculating the overall percent of men or women who appear treats a photo of a crowd the same as many individual portraits.

Researchers wanted to be sure that the pattern of men appearing twice as often as women was consistent across photos that showed any number of people. To do so, researchers calculated the average rate that men and women appeared across posts. For example, consider a photo that shows two women and two men, another that shows no women and one man, and a third with two women and six men. Taken together, the average share of women is 25%, based on (50% + 0% + 25%) / 3. In the same example, the average share of men is 75%, based on (50% + 100% + 75%) / 3).

Across all 22,342 photos from news outlets, the average share of women depicted in each news image is 33%, while the average share of men shown per news image is 67%. These numbers match up with the results based on the percent of all individuals shown.

Women are shown relatively more in news about entertainment, but never more than men overall

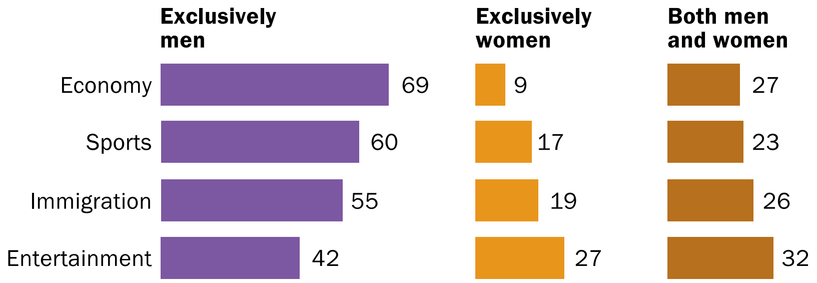
Men appeared more often than women across four categories of news content on Facebook researchers analyzed, but some categories featured relatively more women. Compared with other categories, women are more likely to appear in news photos for posts about entertainment – defined here as news about TV, music or movies – than in news photos for posts that mention the economy, immigration or sports. However, even across these four topics, women *never* appeared more often than men within each group of news images.

The four topics researchers selected include two specific policy issues rated as important by substantial shares of U.S. adults (immigration and the economy) and two broader subjects that people seek news about, sports and entertainment.[3] These four topics are not exhaustive and were selected because they related to areas of public interest and could reliably be identified using computational methods. See the Methodology for more information about how the topics were selected.

Researchers trained a text classification model to determine which news posts related to these four topics. This kind of model is based on human decisions about which words are associated with particular topics in a sample of posts. The model then makes predictions about whether any of the posts mention those topics. All posts with news photos were included when classifying the content of the posts – those that showed people and those that did not.

Women are especially unlikely to appear in news images related to the economy, more likely to appear in images related to TV, music or movies

% of news images with people from 17 national news organizations on Facebook that depict …



Source: Pew Research Center analysis of Facebook news images from 17 national news outlets created April 1–June 30, 2018.
"Men Appear Twice as Often as Women in News Photos on Facebook"

PEW RESEARCH CENTER

The model used text that appeared alongside the post, including its title, caption and comment. Researchers

[3] The first two topics appeared in a survey conducted in June 2018, and the second two appeared in a survey conducted in February and March of the same year.

validated the results of the models by asking human coders to examine a subset of posts.[4] Looking across all 44,056 posts – whether their images showed people or not – the model predicted that 5,678 posts mentioned entertainment, i.e., TV, music or movies (12.9%), 1,296 mentioned economic issues (2.9%), 1,529 mentioned immigration (3.5%) and 1,302 mentioned sports (3.0%).

The gender gap in posts mentioning entertainment was notably smaller than other topic categories. Women appeared much more often in news photos accompanying these posts than in those accompanying posts that mention other topics. The study found that 27% of posts mentioning entertainment exclusively showed women and 42% exclusively showed men. Overall, 58% of posts in this category showed at least one woman while 73% showed at least one man.

By contrast, posts that mentioned the economy showed the largest gender gap among topic categories, and people depicted in those photos were much less likely to show women. In fact, just 9% of these posts exclusively depicted women, while 69% showed exclusively men. The remaining 22% of images showed both men and women, with only 31% of these posts showing at least one woman – less than half the rate that women appeared in posts about entertainment. And 91% of posts that mentioned the economy showed at least one man.

The gender gap within posts about sports was also notable: 40% of the images with identifiable human faces accompanying those posts depicted at least one woman, compared with the 83% that depicted at least one man. Results were similar — but slightly less pronounced – for posts that mentioned immigration.

These differences also appeared when examining the share of people who were women across posts mentioning each topic. In news posts about entertainment, 40% of depicted individuals were women. That number is 22% for news posts about the economy. In news posts about immigration, 33% of individuals were women. For posts about sports, women made up 26% of people shown.[5]

[4] See Methodology for a detailed discussion of topic selection and model validation.

[5] Using the average share measure (which adjusts for the number of people shown in photos by averaging across all photos with people), the rates that women appear are 19% for posts that mention the economy, 26% for posts that mention sports, 31% for posts that mention immigration, and 42% for posts that mention TV, music or movies.

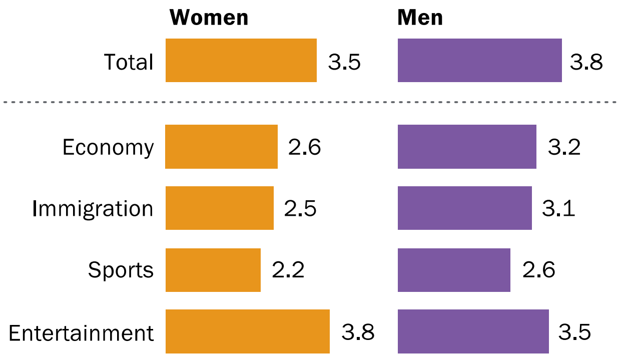
Women's faces appear smaller than men's in news images

When it comes to how prominently individuals' faces were depicted in news photos, there was a modest difference between men and women. Researchers measured the size of women's faces relative to that of men's faces to capture prominence. The technique researchers used to measure faces only captures the size of a person's face, omitting features like hair, jewelry and headwear. The average male face occupied 3.8% of an image on average, while the average female face took up 3.5% of an image. These differences amount to the average male face being shown at a size 10% larger than the average female face.

The average size of women's faces was 19% smaller than the average size of men's faces in posts that discussed the economy. In posts that mentioned entertainment, women's faces were 7% larger than men's faces, on average.

Men's faces appear slightly larger than women's faces, except in photos for posts about TV, music or movies

% of image showing faces of …



Source: Pew Research Center analysis of Facebook news images from 17 national news outlets created April 1-June 30, 2018.
"Men Appear Twice as Often as Women in News Photos on Facebook"

PEW RESEARCH CENTER

Even among news images that featured multiple people, women had the largest face visible in the photograph for only 32% of images.

Across outlets, women never appear more often than men

On average, none of the media outlets examined in the study showed more women than men in news photos. Among the outlets examined here, the percent of women depicted (among all individuals shown) ranged from 25% to 46% per outlet.

Each outlet created 2,592 posts in the three months between April 1 and June 30, 2018, on average, including all posts, whether their photos showed people or not. The outlet level total varied from 878 to 3,975 posts. A total of 1,471 posts – 3.3% of the full sample – were posted by news outlets multiple times with the same text and image; these posts are included in the analysis.

Out of the 17 media outlets, six showed women more prominently than men, on average, in terms of how large their faces appeared.

Acknowledgments

This report is made possible by The Pew Charitable Trusts. It is a collaborative effort based on the input and analysis of the following individuals.

Primary researchers

Onyi Lam, *Computational Social Scientist*
Stefan Wojcik, *Computational Social Scientist*
Adam Hughes, *Computational Social Scientist*
Brian Broderick, *Data Science Engineer*

Research team

Emma Remy, *Data Science Assistant*
Galen Stocking, *Computational Social Scientist*
Patrick van Kessel, *Senior Data Scientist*
Skye Toor, *Data Science Assistant*

Editorial and graphic design

Travis Mitchell, *Copy Editor*
Alissa Scheller, *Information Graphics Designer*

Communications and web publishing

Rachel Weisel, *Communications Manager*
Hannah Klein, *Communications Associate*
Calvin Jordan, *Communications Associate*
Travis Mitchell, *Digital Producer*

In addition, the report benefited from feedback provided by the following Pew Research Center researchers: Claudia Deane, Amy Mitchell, Katerina Matsa, Elisa Shearer, Michael Barthel, Andrew Mercer, Dennis Quinn, Peter Bell, Lee Rainie, Aaron Smith, Kim Parker, Ruth Igielnik and Juliana Horowitz.

Methodology

Terminology

Precision and recall are statistics used to quantify the performance of statistical models making predictions. Low values for precision signify that the model is making a positive prediction about a post or image when in fact the prediction should be negative. Low values of recall signify that the model is systematically missing positive cases that ought to be labeled as such.

Cohen's Kappa is a statistic used to assess interrater reliability. It ranges from 0 (meaning that two separate sets of decisions are only related to each other according to chance) to 1 (meaning that two separate decisions perfectly agree, even adjusting for chance agreement).

Deep learning is a class of machine learning models that is inspired by how biological nervous systems process information. These kinds of models include multiple layers of information that help make predictions. In this report, researchers used deep learning models to predict whether human faces belonged to men or women.

Support vector machines refers to a common machine learning algorithm. The algorithm uses the decision of multiple models and aims to achieve clear separation between classes of data, or predictions. In this case, researchers used support vector machines to classify whether or not posts mentioned particular news topics.

News outlet selection

The study was based on American news outlets whose websites:

1. Have a monthly average of more than 20 million unique visitors from July-September of 2018, according to Comscore's data (Comscore Media Metrix Multi-platform, unique visitors, July-September 2018).

2. Provide original reporting and news and information content for a general audience.

3. Cover a variety of topics rather than specializing in a particular topic (such as weather, sports, politics, business or entertainment)

4. Are based in the U.S.

5. Focus at least in part on national issues (rather than focusing solely on local issues).

After applying these rules, researchers included the following outlets: ABC News, BuzzFeed News, CBS News, CNN, Fox News, HuffPost, NBC News, The New York Times, Newsweek, NPR, Time, U.S. News & World Report, USA Today, The Washington Post, Yahoo News, Vice and Vox.

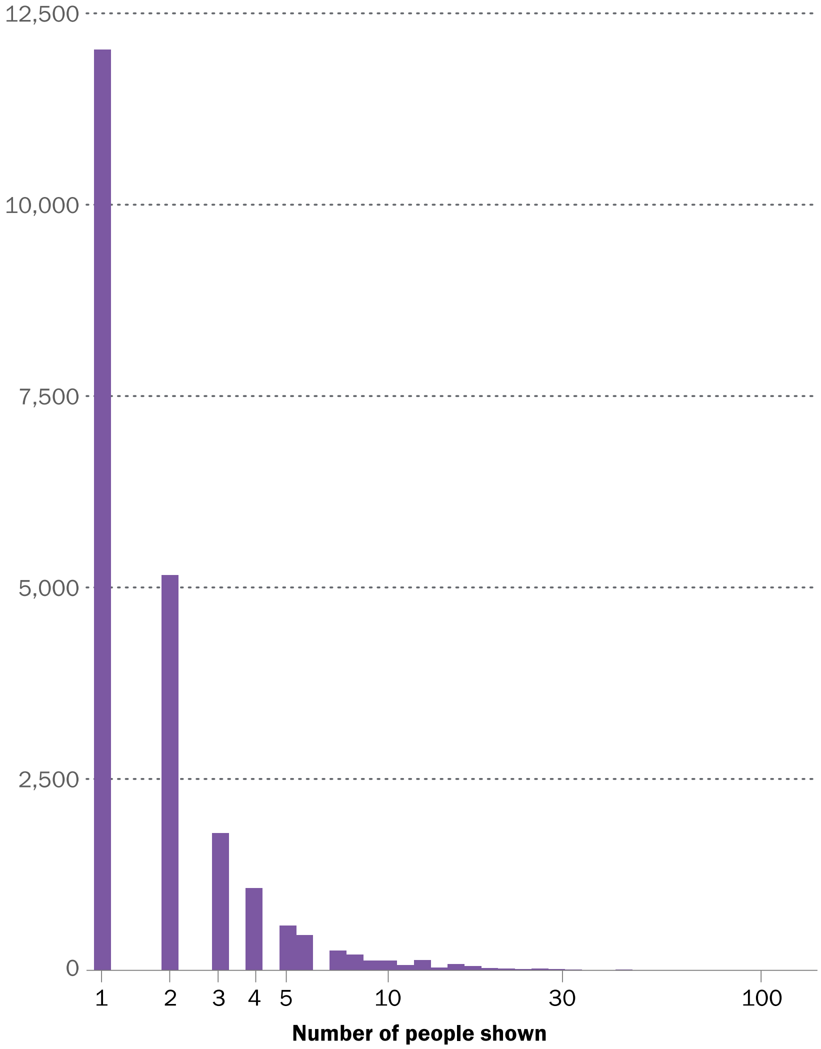
Data collection

To create the dataset used for both analyses, researchers built a data pipeline to streamline image collection, facial recognition and extraction, and facial classification tasks. To ensure that a large number of images could be processed in a timely manner, the team set up a database and analysis environment on the Amazon Web Service (AWS) cloud, which enabled the use of graphics processing units (GPUs) for faster image processing.

Data collection took place in April, May and June of 2018. The information collected about each post includes the title, caption and a brief comment which appeared within the post.

Most Facebook news photos show three or fewer individuals

Number of news images showing __ individuals



Source: Pew Research Center analysis of Facebook news images from 17 national news outlets created April 1-June 30, 2018.
"Men Appear Twice as Much as Women in News Photos on Facebook"

PEW RESEARCH CENTER

Face detection

Researchers used the face detector from the Python library dlib to identify all faces in the image. The program identifies four coordinates of the face: top, right, bottom and left (in pixels). This system achieves 99.4% accuracy on the popular Labeled Faces in the Wild dataset. The research team cropped the faces from the images and stored them as separate files. A total of 44,056 photos were analyzed, 22,342 of which contained identifiable human faces.

Machine vision for gender classification

Researchers used a method called "transfer learning" to train a gender classifier rather than using machine vision methods developed by an outside vendor. In some commercial and noncommercial alternative classifiers, "multitask" learning methods are used to simultaneously perform face detection, landmark localization, pose estimation, gender recognition and other face analysis tasks. The research team's classifier achieved high accuracy for the gender classification task while allowing the research team to monitor a variety of important performance metrics.

Gender classification model training

Recently, research has provided evidence of algorithmic bias in image classification systems from a variety of high profile vendors. This problem is believed to stem from imbalanced training data that often overrepresents white men. For this analysis, researchers decided to train a new gender classification model using a more balanced image training set. However, training an image classifier is a daunting task because collecting a large labeled dataset for training is very time and labor intensive and often is too computationally intensive to actually execute.

To avoid these challenges, the research team relied on a technique called "transfer learning," which involves recycling large pretrained neural networks (a popular class of machine learning models) for more specific classification tasks. The key innovation of this technique is that lower layers of the pretrained neural networks often contain features that are useful across different image classification tasks. Researchers can reuse these pretrained lower layers and fine-tune the top layers for their specific application – in this case, the gender classification task.

The specific pretrained network researchers used is VGG16, implemented in the popular deep learning Python package Keras. The VGG network architecture was introduced by Karen Simonyan and Andrew Zisserman in their 2014 paper "Very Deep Convolutional Networks for Large Scale Image Recognition." The model is trained using ImageNet, which has over 1.2 million images and 1,000 object categories. Other common pretrained models include ResNet and Inception. VGG16 contains 16 weight layers that include several convolution and fully connected layers. The VGG16 network has achieved a 90% top-5 accuracy in ImageNet classification.

Researchers began with the classic architecture of the VGG16 neural network as a base and then added one fully connected layer, one dropout layer and one output layer. The team conducted two rounds of training – one for the layers added for the gender classification task (the custom model), and subsequently one for the upper layers of the VGG base model.

Researchers froze the VGG base weights so that they could not be updated during the first round of training and restricted training during this phase to the custom layers. This choice reflects the fact that weights for the new layers are randomly initialized, so if the VGG weights are allowed to be updated it would destroy the information contained within them. After 20 epochs of training on just the custom model, the team unfroze four top layers of the VGG base and began a second round of training. For the second round of training, researchers implemented an early stopping function. Early stopping checks the progress of the model loss (or error rate) during training and halts training when validation loss value ceases to improve. This serves as both a timesaver and keeps the model from overfitting to the training data.

In order to prevent the model from overfitting to the training images, researchers randomly augmented each image during the training process. These random augmentations included rotations, shifting of the center of the image, zooming in/out, and shearing the image. As such, the model never saw the same image twice during training.

Selecting training images

Image classification systems, even those that draw on pretrained models, require a substantial amount of training and validation data. These systems also demand diverse training samples if they are to be accurate across demographic groups. Researchers took a variety of steps to ensure that the model was accurate when classifying the gender of people from diverse backgrounds.

First, the team located existing datasets used by researchers for image analysis. These include the "Labeled Faces in the Wild" (LFW) and "Bainbridge 10K U.S. Adult Faces" datasets. Second, the team downloaded images of Brazilian politicians from a site that hosts municipal-level election results. Brazil is a racially diverse country, and that is reflected in the demographic diversity in its politicians. Third, researchers created original lists of celebrities who belong to different minority groups and collected 100 images for each individual.

The list of minority celebrities focused on famous black and Asian individuals. The list of famous blacks includes 22 individuals: 11 men and 11 women. The list of famous Asians includes 30 individuals: 15 men and 15 women. Researchers then compiled a list of the most populous 100 countries and downloaded up to 100 images of men and women for each nation-gender combination, respectively (for example, “French man”). This choice helped ensure that the training data included images that feature people from a diverse set of countries, balancing out the overrepresentation of white people in the training dataset. Finally, researchers supplemented this list with a set of 21 celebrity seniors (11 men and 10 women) to help improve model accuracy on older individuals. This allowed researchers to easily build up a demographically diverse dataset of faces with known gender and racial profiles.

Some images feature multiple people. To ensure that the images were directly relevant, a member of the research team reviewed each face in the training datasets manually and removed irrelevant or erroneous faces (e.g., men in images with women). Researchers also removed images that were too blurry, too small and those where much of the face was obscured. In summary, the training data consist of 14,351 men and 12,630 women in images. The images belong to seven different datasets.

Training datasets

Dataset	Number of male faces	Number of female faces	Total
Bainbridge	1,023	753	1,776
Brazil politicians	1,612	1,627	3,239
Labeled Faces in the Wild	2,839	776	3,615
Famous black Americans	755	741	1,496
Famous Asians	796	755	1,551
Country-gender image search	6,629	7,335	13,964
Famous seniors	697	643	1,340

PEW RESEARCH CENTER

Model performance statistics

Data source	Pos. predicted value	Error rate	True positive rate	False positive rate
Bainbridge	0.978	0.022	0.962	0.018
Brazil politicians	0.997	0.003	0.893	0.003
Labeled Faces in the Wild	0.939	0.061	0.953	0.066
Famous black Americans	0.960	0.040	0.966	0.040
Famous Asians	0.943	0.057	0.948	0.053
Country-gender image search	0.899	0.101	0.869	0.029
Famous seniors	0.964	0.036	0.957	0.035

PEW RESEARCH CENTER

Gender classification model performance

To evaluate whether the model was accurate, researchers applied it to a subset of the dataset equivalent to 20% of the image sources: a “held out” set which was not used for training purposes. The model achieved an overall accuracy of 95% on this set of validation data. The model was also accurate on particular subsets of the data, achieving 0.96 positive predictive value on the black celebrities subset, for example.

As a final validation exercise, researchers used an online labor market to create a hand coded random sample of 998 faces. This random subset of images overrepresented men — 629 of the images were coded as male by Mechanical Turk (MTurk) coders. Each face was coded by three online workers. For the 920 faces that had consensus across the three coders, the overall accuracy of this sample is 87%. Using the value 1 for "male" and 0 for "female," the precision and recall of the model were 0.89 and 0.92, respectively, indicating that performance was balanced for both predictions.

Text classification and model performance

To determine whether news posts mentioned particular topics, researchers used a semi-supervised text classification algorithm. The topics were selected because they appeared in contemporaneous Pew Research Center surveys of U.S. adults either as among the most important problems facing the nation (health care, the economy and immigration) or as topics that individuals seek news about (sports and entertainment). Researchers developed a list of keywords related to each topic as "seed words" that initially classified posts as related to each of the topics or not. To narrow down the possible keywords for the entertainment category, researchers operationalized the concept as news mentioning TV, music or movies.

These initial positive cases were used as a training data set, which the researchers then used to fit a support vector machines (SVM) model. The SVM model detects words that co-occur with the "seed words" and uses those additional words to predict which posts were likely to be related to the topics of interest. The model also avoids the "seed words" associated with the other topics. The seed words of other topics help the model determine the negative cases. For example, when applying the model for sports, a post might use a seemingly relevant term like "winner" but *also* use terms associated with the economy like "trade war." In such a case, the model is especially unlikely to classify the post as mentioning sports.

To prepare the data needed to train the model, researchers preprocessed the text by removing stop words. These words include commonly used English words such as "and," "the," or "of" that do not provide much information about the content of the text. Researchers then used the TfidfVectorizer in the sklearn python library to convert the text to tokens, including phrases that were one, two or three words long. The model was then applied to the full dataset, resulting in a prediction about whether every post mentioned one of the topics or not.

To validate this approach, researchers selected 1,100 posts for human coding. Since the prevalence of the posts that actually discuss each of the five topics was low, researchers used oversampling – based on model-based estimates – to increase the representation of positive cases in the validation sample. Specifically, researchers randomly selected approximately half of the posts that were tagged as positive by the SVM model, and the other half was tagged as negative. After removing duplicate posts, researchers classified 1,061 posts, determining whether they mentioned any of the five topics of interest. Interrater reliability statistics are weighted to reflect the oversampling process.

Initial text classification validation results

Topic	Precision	Recall	Weighted Kappa (model vs. human)	Weighted Kappa (human vs. human)
Economy	0.78	0.90	0.84	0.70
Immigration	0.84	0.97	0.90	0.78
Health care	0.49	0.84	0.51	0.65
Sports	0.79	0.93	0.81	0.86
TV, music or movies	0.63	0.82	0.68	0.92

PEW RESEARCH CENTER

Two in-house coders classified the same subset of posts (406) to ensure that humans could reasonably agree on whether or not a post mentioned each issue. After conducting the content coding, researchers resolved disagreements and created a consolidated set of human decisions to compare the model against. The vast majority of posts were coded as mentioning to a single topic, 55 posts in total (0.1%) were coded as mentioning to multiple topics.

The performance of the model and the human coders' agreement with each other is described in the table above. The model can be assessed via precision, recall, and weighted kappa, each of which compares how well the model's decisions correspond with those of the human coders. The final column shows the weighted kappa for the subset of posts coded by two coders, comparing their decisions against each other.

Researchers found that the health care topic had low precision, suggesting that many posts that the model identified as mentioning the topic did not in fact mention it. A manual review of posts revealed that the model incorrectly classified posts that mentioned personal health and wellness as mentioning health care, while the coders were focused on health care policy or general health care issues. Since the words associated with health care policy were so similar to those associated with personal health and wellness, researchers decided to exclude this topic from further analysis.

The precision value for TV, music or movies was also low, but it was clear in this case that false positives were decreasing model performance, due to both the "seed" keywords associated with the topic and the penalty that the model applied to mislabeled posts. In response, researchers revised the keyword list and also adjusted the model parameter that controls the size of the penalty assigned to mislabeled posts. If that model parameter is smaller, the model will achieve better separation between the positive and negative posts. So researchers lowered the parameter from 1 to 0.05.

Since changing the model risks overfitting to the data, researchers separately drew a new sample of 100 posts (50 positive and 50 negative) to conduct model validation. Two in-house coders classified these posts. After reevaluating the models, researchers arrived at the following performance statistics. [6]

Final text classification validation results

Topic	Precision	Recall	Weighted Kappa (model vs. human)	Weighted Kappa (human vs. human)
Economy	0.83	0.83	0.80	0.71
Immigration	0.85	0.96	0.89	0.79
Sports	0.86	0.91	0.86	0.87
TV, music or movies	0.78	0.95	0.72	0.94

PEW RESEARCH CENTER

Performance statistics for the topic including TV, music or movies was substantially better in this round of validation: Both precision and recall increased. The model performance of the other topics also changed slightly due to the fact that seed words for one topic identify negative cases for other topics; since the keyword list for TV, music or movies changed in this round, it also affected the results of the other topics.

[6] Note that the coder to coder weighted kappa for TV, music, or movies is based on a revised validation set of 100 oversampled and double-coded posts.