CIS399: The Science of Data Ethics **Lecture 15: Mar 14, 2019**
Instructor: Michael Kearns & Ani Nenkova **Scribes: Benjamin Hoover and Alaina Silverman**

1 Recap - Autocomplete

What are appropriate autocorrect results for an individual?

Sample results taken from Google autocomplete:
Louis von Ahm - ”net worth”, ”ted talk”, ”married”, ”Carnegie Mellon”
Tuuli Lappalamen - ”wedding”, ”cv”, ”husband”, ”twitter”, ”Google scholar”
Are these types of results - very personal information queries - acceptable?

2 Image Processing and Crowdsourced Data

Early image recognition initially used words on the page and in urls as a proxy for what an image represents. To create more advanced algorithms, the collection of user generated labels were needed.

We investigated how crowd-sourced data has been collected and the ethical concerns of the methods.

2.1 Mechanical Turk Crowd-sourcing

Mechanical Turk allows companies/researchers an easy way to outsource recognition/classification tasks, but there are ethical concerns about its usage.

1. How much is appropriate to pay workers? A dollar an hour can be considered a lot in some parts of the world.

   In papers it is expected to say how much crowd-sourced workers were paid; usually it is expected to pay at least minimum wage.

2. Companies may put up private data up for a crowd-sourced data labeling task, raising questions about data privacy

3. The use of Mechanical Turk for behavioral experiments, as opposed to an academic environment, to answer questions such as:

   - If you pay people more, will they do better/longer work on Mechanical Turk?

   could be controversial as workers may find out that they were paid different amounts for the same task.

2.2 The ESP Game

Luis von Ahn and Laura Babbage developed the ESP game in 2003 as a way to crowdsource image recognition tasks.

Idea: can we trick people into having fun while actually having them perform useful classification tasks?

In the game, an image was drawn, and people had to type word to classify image, and got more points if other people put the same thing.

Was successful, some people played 20 hours/ week, 30,000 people played, over 4 million labels for 1 million images were collected.

The resulting dataset was bought be Google.

Questions:

Is it ok to "trick" people into challenges/tasks that are actually meant to collect data?

Is the Facebook "10 year challenge" actually a data collection task?

Are Facebook games collecting user's information for research tasks?

2.3 Neural Networks

The advent of Neural Networks combined with pre-existing datasets has led to impressive accuracy on image classifiers.

For example, see the Google Photos app's ability to classify photos based on subject.

2.4 ImageNet

Developed by Princeton, 1.2 million labeled images, with categories, became a standard benchmark for image recognition.

However, it still flaws in that it underrepresnts many regions and cultures. See section 3.1.

3 Unfairness in Models

3.1 Representation Bias in Image Databases

Refer to Suresh et. al. study ("A Framework for Understanding Unintended Consequences of Machine Learning")

45% of images were taken in US

1% of images were taken in China

2.1% of images were taken in India

Majority of remaining images are from elsewhere in North America or in Western Europe.

This makes it difficult for the system to classify things that only exist in other (non Western) cultures. It can also lead to issues where things are represented differently in different cultures, such weddings, which are quite different in America vs India.

3.2 Intersectional Fairness and Gerrymandering

Oftentimes fairness goals are set and achieved for individual groups or features, such as gender and race. However, the intersection of such features are not guaranteed fairness. This means that the treatment of individuals who have multiple targeted features is not used as a benchmark.

When intersectional fairness is not achieved, fairness gerrymandering occurs. This gerrymandering is the appearance of fairness without true fairness; an concise example is a company that hires black men and white women, but no black women.

3.3 Race and Gender in Recognition

We have seen that people recognition models are trained on data-sets comprised of mostly white men with pale complexions, with women and people of other races and complexions being a small part of the data.

In gender recognition studies this has resulted in near perfect results on assigning gender to white men. Key statistics are an 8 percent better performance on men than women, and 10-20 percent better performance on lighter skinned individuals than on darker skinned individuals.

Additionally, autonomous cars are less likely to recognize people with darker skin as pedestrians, resulting in more accidents involving those individuals. There are also anecdotes of products not working as well for people with darker skin - for example, Xbox tracking and soap dispensers.

This is all evidence for improper performance benchmarking and the fact that poor data has significant impact on the performance of models on specific groups.

3.4 Proposed Solutions

There were several solutions proposed in class:

1) Thinking about the data more. There need to be industry standards that are met before a data-set is released. Data collection and how representational it is are both important.

2) Models should broadcast what training data was used.

3) Modes should be judged by many separate benchmarks, and especially use separate benchmarks for intersectional groups.

4 African American Vernacular English (AAVE) in Tweet Data

The locations that users tweet from are used to create 'soft labels' on the racial profile of the tweeter and the type of language used in the tweets. This means that a tweet sent out from a predominantly black neighborhood will be labeled as using AAVE, with the same process for labeling tweets with Standard English.

It was found that tweets written in AAVE were less likely to be recognized as English as all by data scraping models, especially the shorter tweets. A subsequent effect of this is that these AAVE tweets are not used for analysis.

Additionally, automated rudeness flagging systems often flag AAVE tweets as rude or hostile regardless of their actual content.

π
Happy Pi Day!