Title of the Pitch

Bias Detection in Unstructured Datasets for Auto AI

Pitch Owner(s):

Balaji Ganesan, Kalapriya Kannan

Problem:

What industry is this problem in?

Cross industry

Summary of the problem you want to solve: describe what real world problem needs to be solved

Given any unstructured dataset, identify any skew in the dataset for protected classes (sensitive personal data entity types like name, gender, age, ethnicity).

To that end, we need to be able to identify and extract personal data entities from unstructured text. We already have a research asset that accomplishes this.

https://data-discovery-demo.mybluemix.net/

We then have to determine if the distribution of such entities is biased towards a subset of possible types. We are proposing a solution that uses benchmark tasks, flipping of data and a combination of generative and discriminative models to determine if the dataset is biased on any of the protected classes. Please see detailed description of the solution later in the 'Solution' section below.

Who has the problem and how do you know it is serious?

Clients using Watson Studio and Watson OpenScale may be inadvertently building and deploying models that are biased, because their training data is not diverse enough. This skew in the training data will lead to sub-optimal performance on unseen data, but may also result in a biased model and expose clients to adverse publicity (see examples below). The issue is serious enough for IBM to release a dataset of human face images and AIF360 toolkit.

This problem is lot more pronounced as we approach the Auto AI and Auto Data goals, where the clients will rely on IBM to handle the data pipeline, feature generation and model building at scale.

Examples of bias in the real world from dataset issues:

1. Parole violation prediction

In a model predicting whether a person is likely to violate parole conditions, and hence determine if a person should be granted parole, the location and ethnicity of people was an important factor. It was found that people from a particular race, or a neighborhood that sees more crimes and parole violations, may at the macro level continue to see more parole violations. So the training data reflects that. While in Machine Learning, this might be considered a “good feature” to predict parole violation during testing, such a model will be considered biased towards an individual. Such models will need more scrutiny from ethics point of view. Our solution will highlight any such skew in datasets and make the model builders aware of potential bias in the model prediction.

2. Deep Learning models predicting on irrelevant features.

In a model built to predict the occurrence of malevolent cancer in patients, a model was performing very well. But upon closer look, and much to the horror of the model builders, the model was learning the location of the hospital (printed on radiology reports) as a feature. The model builders did not mean to use that as a feature, and infact it was the Neural model that was coming up with features. Our solution will highlight occurrence of such variables in the dataset and help eliminate them from consideration.

How do people deal with this problem today?

For the large part, people are unaware of the biases in their unstructured training datasets. This problem might eventually get handled as an accuracy problem where wrong model predictions are attributed to the other kind of “bias” i.e. overfitting problem in Machine Learning.

There are solutions to test a model for bias by using adversarial solutions, flipping features etc, especially in structured data. We’re proposing to solve the problem much ahead in the AI cycle, by analyzing the datasets for lack of diversity on sensitive personal data entities.

Solution:

Summary of the solution (optional 1 image)

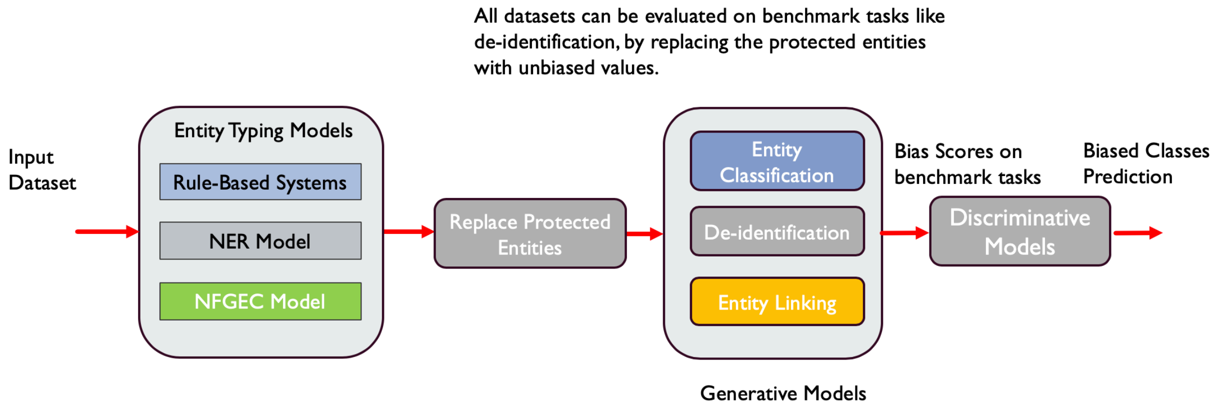
The solution we are proposing involves the following steps.

1. Identify person data entities in the unstructured training data (independent of the downstream task). Already done by us.
2. Replace the identified personal data entities with known unbiased entities (drawn from a vocabulary for each protected class where the entities are normally distributed).

3. Now use the input data (with replaced entities) on benchmark tasks. These models act as generative models for predicting the bias with respect to one or more of the protected classes.
4. Finally, a discriminative model (a classifier), determines if the input data is biased on any of the protected classes.

The above solution is captured in the architecture diagram below. The input to this pipeline will be unstructured data, preferably as sentences. The output will be a score for each of the personal data classes, indicating if the dataset is biased with respect to that class.

Bias Detection in Unstructured Datasets – Solution Diagram


All datasets can be evaluated on benchmark tasks like de-identification, by replacing the protected entities with unbiased values.
Input Dataset
Entity Typing Models
Rule-Based Systems
NER Model
NFGEC Model
Replace Protected Entities
Entity Classification
De-identification
Entity Linking
Generative Models
Bias Scores on benchmark tasks
Biased Classes Prediction
Discriminative Models

Connection to other pitches

We’re not aware of other pitches yet, but we’ve worked on 2 projects related to this proposal.

- Data Discovery in Data lakes

- Joint Learning of Mention Detection and Entity Classification for Medical Text De-identification.

Desirability:

- Why is this solution better than the alternatives?

  Detecting bias in training data is lot cheaper than debugging model issues in downstream applications arising out of bias.

- Why would someone choose this over what exists already?

We’re not aware of publicly available solutions in this space. IBM is a pioneer in this space with AIF360 and other efforts.

Strategic Alignment:

- What is IBM's competitive advantage?

  We have first mover advantage in bias detection. We have enterprise customers who may be interested in finding biases in their data, but without actually exposing the data to others, including IBM. The ability to automatically detect bias with minimal human intervention is particularly useful for customers of IBM Private Cloud and IBM Hybrid Cloud.

- How does this connect to other parts of IBM's business?

  Any business unit that builds models for client projects, or incorporates machine learning models in their product, can use this work to assess if their training data is biased.

Technical Feasibility

- What evidence do we have that this is technically feasible? (e.g. prior challenges, demos, papers, working code)

1. IBM through AIF360 has already had some success detecting bias in structured data. We’re familiar with the techniques in that effort and are building on them for unstructured data.

2. We’ve worked on identification of personal data entities under two AI challenges.

   Our two manuscripts on Entity Classification are now available on arxiv.

   - Riddhiman Dasgupta, Balaji Ganesan, Aswin Kannan, Berthold Reinwald, and Arun Kumar. "Fine Grained Classification of Personal Data Entities." *arXiv preprint arXiv:1811.09368* (2018). https://arxiv.org/abs/1811.09368

   - Abishek, Amar Prakash Azad, Balaji Ganesan, Ashish Anand, and Amit Awekar. "A Unified Labeling Approach by Pooling Diverse Datasets for Entity Typing." *arXiv preprint arXiv:1810.08782* (2018). https://arxiv.org/abs/1810.08782

3. We organized a hackathon with IIM Ahmedabad (a leading business school in India) where the task is to impute personal data entities in already redacted text. We're organized a similar challenge at AMLD 2020. We're working on our own solution to this problem.

Viability

- What parts of this solution are differentiating for IBM, i.e. hard for a competitor to replicate?

  With large number of models trained on client data in private and hybrid clouds, this problem is unique to IBM and few of our competitors. But this is also an opportunity to come up with a bias detection solution, get feedback and scale to a large number of datasets. The generative models that we propose to learn will need access to diverse datasets available in an industrial lab, which is hard to replicate by many of our customers.

- How crowded or competitive is the market?

  While there are a number of academic efforts, and presumably copyrighted solutions in few companies, we are not aware of any publicly available competitive products in this space.