Fairness and Transparency for Big Data Analysis

Gautam Kishore Shahi
shahi@l3s.de

Content

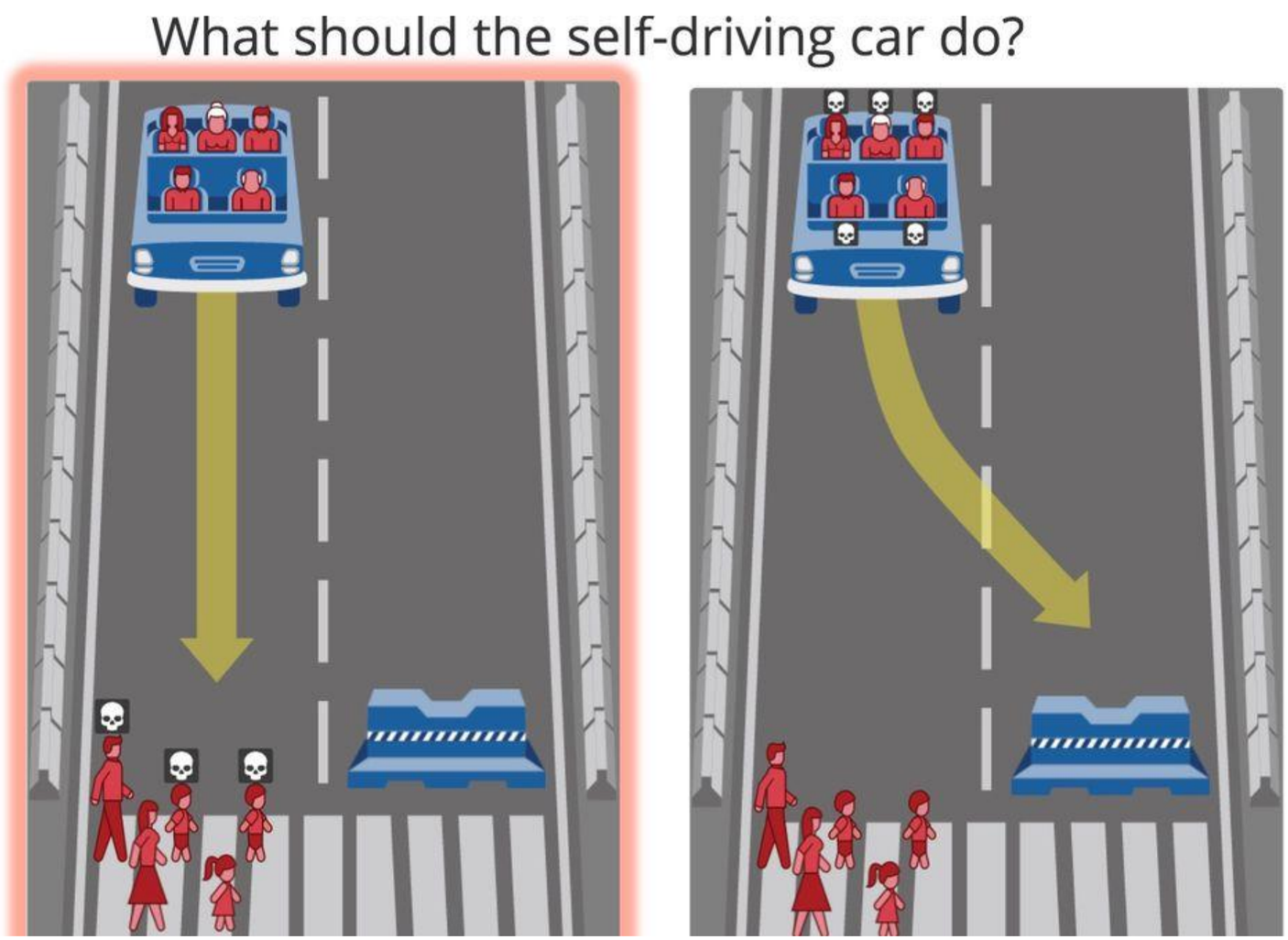
- Introduction
- Motivation
- Paper
  - Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity
  - Semantics derived automatically from language corpora necessarily contain human biases
- Conclusion and Discussion
- References

Introduction

- Fairness
  - Fairness is not a technological problem but unfair behavior can be replicated/automated using technology.
    - Discrimination
      - Gender, Ethnicity etc.

- Transparency
  - Transparency implies openness, communication, and accountability.
  - Understanding why a ML Model Provides a Given Prediction
    - Decision on Credit Card

Example of an Bias



Transparency

Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity

- Introduction
- Research Question
- Approach & Findings
- Data set
- Methodology
- Results
- Discussion & Limitation

Introduction

- Longitudinal Impact of a collaborative filter based recommender system
- Amazon once reported 35% in sales came from recommender system
- Sales of Netflix increased by 75% in 2012
- Filter bubble effects on the longitudinal data for the recommender system

Research Question

- Do recommender system expose users to narrower content over time?
- How does the experiences of user who take recommendation differ from that of users who do not regularly take recommendations?

Approach & Findings

Approach

- Seperate rate users into categories based on how often they consume recommend content
  - Eg.- People don’t follow the recommendation system
- Metric exploring changes in the diversity of consumed items over time.

Findings

- Novet set of method to study the effect of recommendation on the users
- Quantitative evidence for users who takes recommender system or who don’t uses it.
- Top recommended item became more similar.

Dataset

- Movielens Data
- 217267 unique user
- 20 million movie ratings for 20000 movies
- Longitudinal Data

Methods

- Following Group- 286 Users
- Ignoring Group - 430 Users

Measuring Content Diversity

$$d_{(m_i,m_j)} = \sqrt{\sum_{k=1}^{m}[rel(t_k, m_i) - rel(t_k, m_j)]^2}$$

Results

- Do recommender system expose users to narrower content over time?

  Compared the content diversity at the beginning and at the end of user's observed rating history.

	At the beginning	**At the end**	**Within Group p-value**
All User	25.02	24.67	2.43e-06
Following Group	25.22	24.80	0.014
ignoring Group	24.74	24.52	0.087
Between group p value	0.0037	0,0406	

Results

- How does the experiences of user who take recommendation differ from that of users who do not regularly take recommendations?
  - Does taking recommendation lowe the summed content diversity?
    - There is no difference between the content diversity of the consumed movies by the group
    - Following group watch more diverse movies than ignoring group
  - Did the following group have better experience?
    - Ignoring Group have better experiences
    - Ignoring group watch less enjoyable movies
  - What does the changes of rating average mean?
    - First rating, average rating of 3.63 at 58.93th percentile for following group while 3.74 rating observed in 63.63th group
    - While in last rating ignoring group has 44.77th % and 1.21 drop in following group

Discussion & Limitation

Discussion

- Following group gets more diverse movie due to personalization
- Following group narrowed the content diversity of the rated movies

Limitation

- Follower of movielens?
- Restricted to use top picks for you

Semantics derived automatically from language corpora necessarily contain human biases

- Introduction
- Contributions
- Data
- Methodology
- Results
- Discussion

Introduction

- Bias in Humans and Machines
  - In AI bias is prior information, a necessary prerequisite for intelligence.
  - Harmful bias is called prejudice.
- Bias and prejudice in AI termed word embeddings
- Implicit Association Test(IAT)

Contributions

- Demonstrating bias and prejudice in text
- Measured the distance between the vector
  - Programmer is closer to the man than woman
- Like, IAT, don’t compare the word
- IAT applies to the individual human subjects while embedding derived from the aggregate writings of human on the web.

Dataset

- Glove word embedding
  - 840 Bullions tokens
  - 300 dimensions

Method

- **Word Embedding Association Test(WEAT)**
  Target word- Programmer and nurse
  Attribute- Man, Woman

$$s(X,Y,A,B) = \sum_{x \in X} s(x,A,B) - \sum_{y \in Y} s(y,A,B)$$

$$s(w,A,B) = \text{mean}_{a \in A} \cos(\vec{w},\vec{a}) - \text{mean}_{b \in B} \cos(\vec{w},\vec{b})$$

$$\Pr_i[s(X_i,Y_i,A,B) > s(X,Y,A,B)]$$

$$\frac{\text{mean}_{x \in X} s(x,A,B) - \text{mean}_{y \in Y} s(y,A,B)}{\text{std-dev}_{w \in X \cup Y} s(w,A,B)}$$

Method

- **Word Embedding Factual Association Test (WEFAT)**
  For a single set with two difference set of attribute A and B

$$s(w,A,B) = \frac{\text{mean}_{a\in A}\cos(\vec{w},\vec{a}) - \text{mean}_{b\in B}\cos(\vec{w},\vec{b})}{\text{std-dev}_{x\in A\cup B}\cos(\vec{w},\vec{x})}$$

The null hypothesis is that there is no association between s(w,A,B) and pw. We test the null hypothesis using a linear regression analysis to predict the latter from the former.

WEAT and WEFAT

- The subject in IAT test became words not people.
- IAT can measure differential association a single of target and concept and an attribute but WEAT needs a pair.
- WEFAT
  - Occupation word vectors embed knowledge of the gender.
  - Test if androgynous names embed knowledge of how often the name is given to boys and girl.

Results

Racial Bias

- Original finding

  European American was found to be easier to associate with pleasant (health, love, peace) than unpleasant(abuse, crash) terms compare to the bundle of African American names. Effect size of 1.17 and p-value of 10-6

- Author's finding

In the corpus the original African American names with sufficient frequency was not there like Tawanda, Effect size 1.41 and p-value of 10-8

Results

Gender Bias

- Original finding

  39797 intepretable subjects, female names were career words of Effect size 0.72 and p-value of 10-2

- Author's finding

  female names are more associated with family and male with career words, Effect size 1.91 and p-value of 10-3

Discussion

- Consequences of bias in Humans and Machines
  - Individual expectations, public policy and even law.
- Effects of Bias in NLP
  - European-American name will have a higher sentiment score than African-American name
  - Machine Translation

Challenges in addressing Bias

- Word don’t merely pickup specific bias but rather entire spectrum of human biases reflected in language
- Who will correct the bias?
- Biases result extant as well as historic inequalities in the world

Conclusion and Discussion

- Awareness is better than blindness
- Choose the right learning model for the problem.
- Choose a representative training data set
- Monitor performance using real data.

Refrences

- https://en.wikipedia.org/wiki/Transparency(behavior)
- https://www.theverge.com/2018/10/24/18013392/self-driving-car-ethics-dilemma-mit-study-moral-machine-results
- https://iamcheated.indianmoney.com/blogs/reasons-why-credit-card-application-gets-rejected
- Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, Joseph A. Konstan. *Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity*. WWW '14.
- Aylin Caliskan-Islam, Joanna J. Bryson, Arvind Narayanan. Semantics derived automatically from language corpora necessarily contain human biases. 2016.

Discussion