Fair Two-Sided Spam Detection Using Machine Learning

Kenny Kwock, Jiaxin Liu, Sihong Xie

CSE Department, Lehigh University, Bethlehem PA

Abstract

Online product review aggregating websites such as Yelp and Amazon, allows users to leave reviews based on their level of satisfaction on a specific product. However, the opportunity for users to leave reviews introduces the possibility for spammer reviews, i.e., reviews that serve to hurt or promote a business, disregarding their actual experience with the product. Utilizing Graphical Neural Networks, and Multi-objective Optimization, we construct a graphical model of the 3 different datasets, to train the model to multi-task: detect spammers and reduce discrimination and bias toward protected products.

Problem Definition

- Model data as undirected graph, three different nodes: User, Review, Product.
- Every node contains array of features and ground-truth label to identify it as a spammer or non-spammer.
- Input data into Neural Network, each hidden layer is a Graphical Convolutional Layer (or known as GCN for short)
- Output data is a vector of predicted spammer and non-spammer
- Model attempts to optimizes parameters based on spammer detection accuracy, and discrimination bias

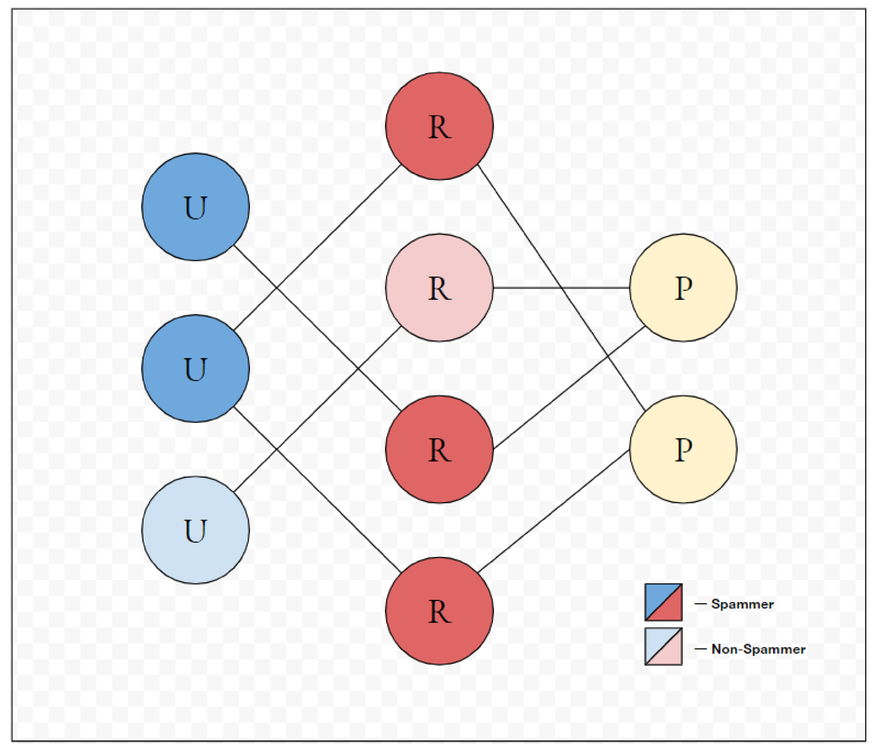
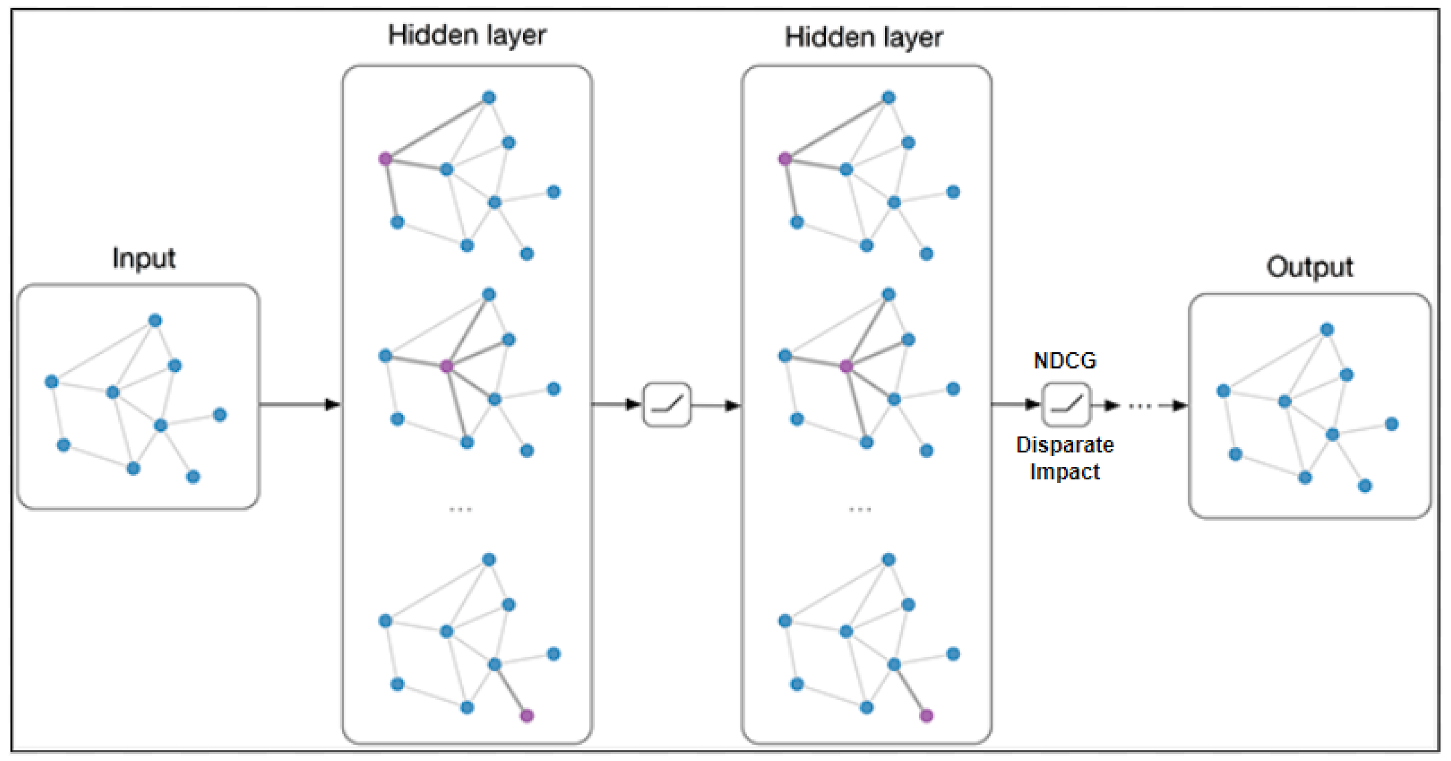


Figure 1.1

Figure 1.2



Methodology

The NN will train under 2 objective functions: NDCG and Disparate Impact

Objective 1: Normalized Discounted Cumulative Gain (NDCG)

Rank	Ideal Rank	Nodes	Influence	Predicted
1	1	Spammer	1.00	0.98
2	2	Spammer	1.00	0.92
3	5	Non-Spammer	0.00	0.85
4	3	Spammer	1.00	0.77
5	6	Non-Spammer	0.00	0.51
6	4	Spammer	1.00	0.39
7	7	Non-Spammer	0.00	0.20
8	8	Non-Spammer	0.00	0.12
9	9	Non-Spammer	0.00	0.05

Figure 2.1

- The **NDCG metric** measures how well the model detects spammers on a ranking scale
- Ideal score is achieved when the model ranks all spammers in the top, and non-spammers in the bottom
- Model DCG score is calculated based on how model ranks each review, NDCG score compares model DCG score and ideal DCG score
- Objective is to maximize this value, as close to 1 as possible

$$\mathrm{DCG}_n = \sum_{i=1}^{n} \frac{g(i)}{\log_2(R_i + 1)}$$

Equation 2.1 $g(i) \in \{0, 1\}, R_i \in \{1, 2, \cdots, n\}$

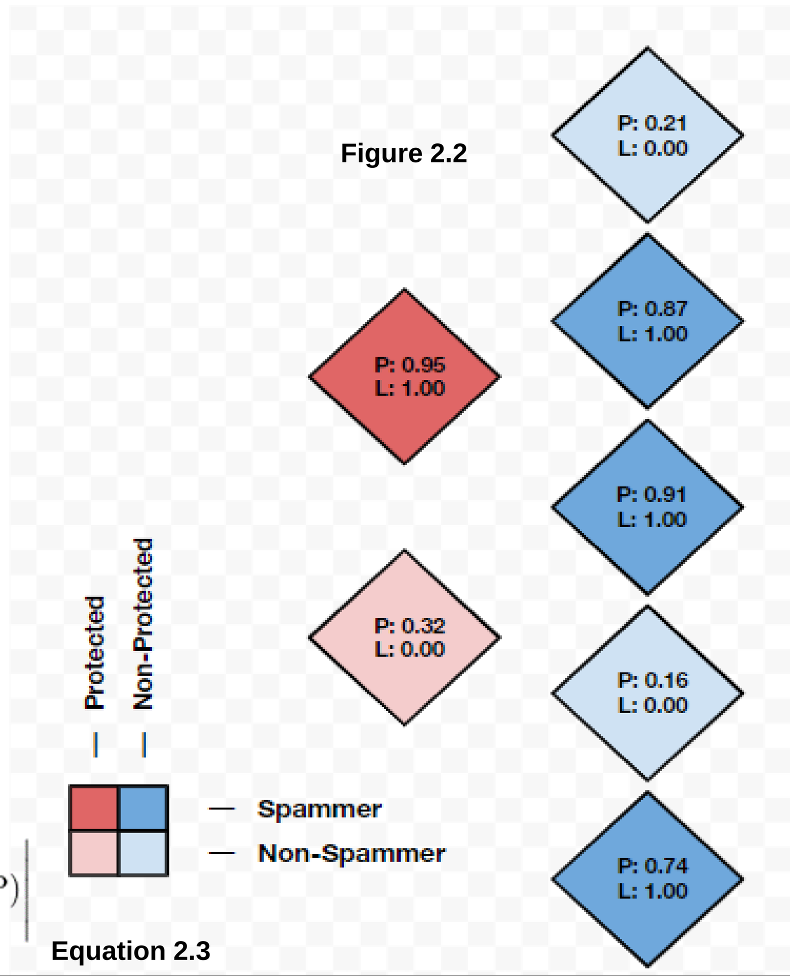
$$\mathrm{nDCG} = \frac{\mathrm{DCG}_{\mathrm{predict}}}{\mathrm{DCG}_{\mathrm{ideal}}} \in [0, 1]$$

Equation 2.2

Objective 2: Disparate Impact

- The **Disparate Impact** metric measures the level of discrimination difference between two different groups of reviews
- Ideal score is achieved when the model detects spammers in equal proportion for protected and non-protected groups
- Score is measured based on how much the model prediction, and spammer score differ from each of the two groups
- Objective is to minimize this value, as close to 0 as possible

Figure 2.2



$$\mathrm{Loss} = \left| \frac{1}{|N_{\mathrm{P}}|} \sum_{n \in N_{\mathrm{P}}} Pr(\hat{Y}|Y, \mathrm{P}) - \frac{1}{|N_{\mathrm{NP}}|} \sum_{n \in N_{\mathrm{NP}}} Pr(\hat{Y}|Y, \mathrm{NP}) \right|$$

Equation 2.3

Final Objective: Optimize NDCG and Disparate Impact Simultaneously

- Calculate optimal proportion of spammer detection and fairness bias
- Model optimal final objective as L2-norm of the sum of both objective functions
- In order to optimize the model further, constantly update the neural network parameter based on the update function

$$\text{Final Objective - } \begin{cases} \text{minimize} & \|\sum_{i=1}^{m} \lambda_i \nabla_\theta f_i(x, y)\|_2^2 \\ \text{subject to} & \|\vec{\lambda}\| = 1, \lambda_i \geq 0 \end{cases}$$

$$\theta^{(t+1)} \Leftarrow \theta^{(t)} - \eta_y \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x, y)$$

Equation 3

Experimental Results

- Train 3 different models with 3 different datasets, plot to check convergence

Training Model Results:

Objective 1: **NDCG Score** *(0 – Minimum, 1 – Maximum)*

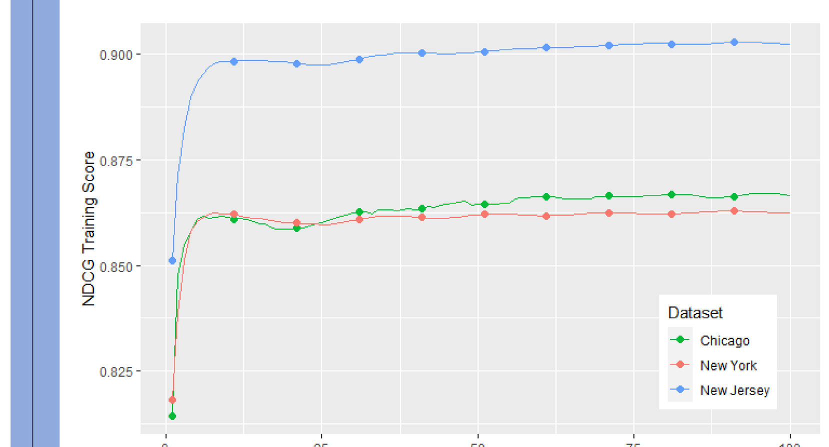


Figure 4.1

Objective 2: **Disparate Impact** *(0 – Maximum, 1 – Minimum)*

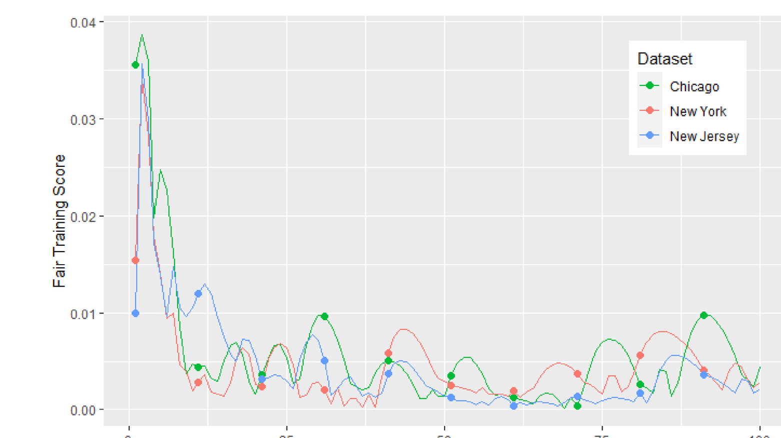


Figure 4.2

- Model trained under 100 iterations of loss calculation and parameter updates
- Both objectives (NDCG, Disparate Impact) converge nicely, however, Disparate Impact results are a bit stochastic
- Visualize lambda multipliers (blue – NDCG, red – Disparate Impact)
- Minimum is achieved when more emphasis is placed on NDCG than Disparate Impact

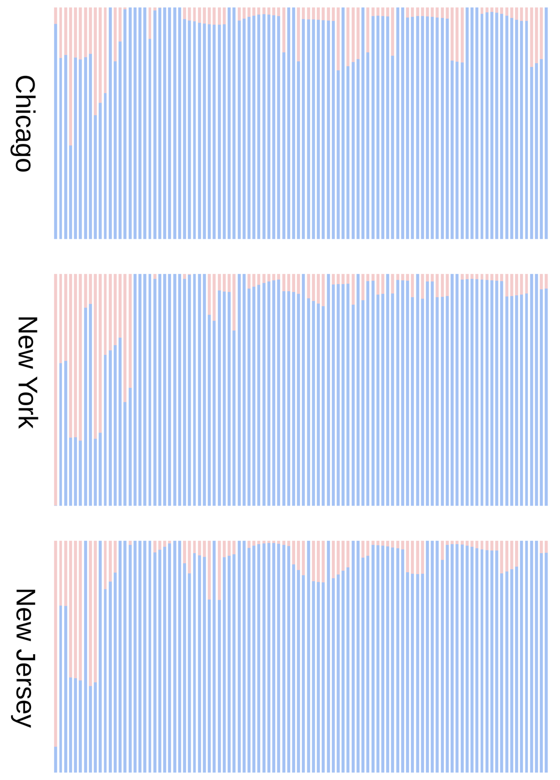


Figure 4.3-4.5

Testing Model Results:

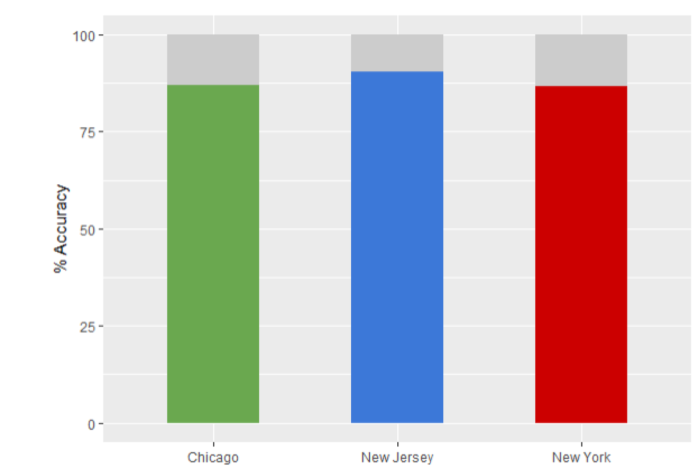
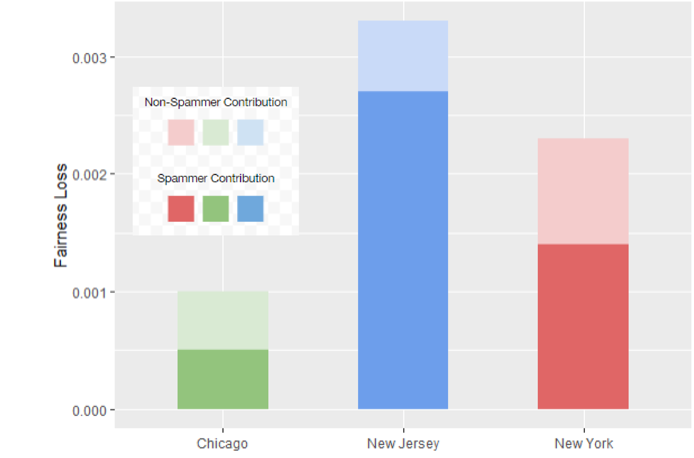


Figure 4.6-4.7



- Testing model, achieves a relatively high accuracy in Spam Detection
- Fairness Loss is high for New Jersey relative to other datasets, but incredibly low in general
- Spammer reviews contribute majority to Fairness Loss
- Model can simultaneously train on two objectives of both Spam Detecting and Fairness/Discrimination Bias

Future Work

- The Multi-Objective Optimization framework works with two objectives
- Applications of this work can be realized when training models to be fairer in other areas of work fairness, discrimination, and more
- Potentially introduce more than 2 objectives in Machine Learning model

Acknowledgements

We thank Sheldon Xu for implementing various fairness metrics.
Jiaxin Liu and Sihong Xie are supported by NSF grants CNS-1931042 and IIS-2008155.
Kenny Kwock and Kai Burkholder are supported by CNS-1757787

The MOO framework is inspired by discussions with Luis N. Vicente from ISE

References: S, Liu, L.N. Vicente, The stochastic multi-gradient algorithm for multi-objective optimization and its application to supervised machine learning, 2021. https://arxiv.org/pdf/1907.04472.pdf