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 an undirected graph, three different nodes: User, Review, Product.
- Each node contains an 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 optimize parameters based on spammer detection accuracy, and discrimination bias.

Methodology

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

Objective 1: Normalized Discounted Cumulative Gain (NDCG)

- 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 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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Ideal Rank</th>
<th>Nodes</th>
<th>Influence</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Spammer</td>
<td>1.00</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Spammer</td>
<td>1.00</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Non-Spammer</td>
<td>0.00</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Spammer</td>
<td>1.00</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Spammer</td>
<td>1.00</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Non-Spammer</td>
<td>0.00</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Non-Spammer</td>
<td>0.00</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Non-Spammer</td>
<td>0.00</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Non-Spammer</td>
<td>0.00</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Non-Spammer</td>
<td>0.00</td>
<td>0.12</td>
<td></td>
</tr>
</tbody>
</table>

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

\[ \text{Fairness Loss} = \frac{1}{|B_1|} \sum_{y \in B_1} \text{Pr}(y | P) - \frac{1}{|B_2|} \sum_{y \in B_2} \text{Pr}(y | N) \]

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

\[ \text{Final Objective} - \min_{\theta} \sum_{i=1}^{m} \lambda_i \| \nabla f_i(x, y) \|_2^2 \]

Experimental Results

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

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 Burkholtzer are supported by CNS-1757787

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