CS-GY 9223 Project Proposal - Visualizing Bias in Sentiment Analysis Models

Jasper Duan
NYU
zd793@nyu.edu

Diana Levy
NYU
dl2666@nyu.edu

1. Introduction

Natural language processing is widely used in an increasingly data-centric industry. Despite being very effective in processing textual data, NLP faces setbacks when applied in practice because it often also learns biases found within text. As a result of algorithmic unfairness, the use of NLP may perpetuate the biases that progressive societies aim to remove. Because bias in data can exist in many shapes and forms and have varying results in downstream tasks, addressing biases in a model is a complex task where different analyses are needed for each type. [1]

Project goal: In this work, we aim to build a system that visualizes the bias in deep neural network for sentiment analysis. Through this system, we hope to enable to user to more effectively discover biases within the dataset and the model. Specifically, we want to facilitate the discovery of not just direct biases (ie gender) but also second-order biases (ie masculine text) [2] in a sentiment analysis model.

2. Related Work

Because bias in machine learning can determine the usability of an application, many companies have a vested interest in developing tools for identifying out bias. The What-if Tool by Google is a generalized tool for analyzing feature impact on a prediction, which can be applied to bias detection when protected groups are labeled features. The AI Fairness 360 by IBM goes one step further by implementing dataset mitigation techniques. For neural networks, the Testing with Concept Activation Vectors (TCAV) by Google is a interpretability method that can measure sensitivity to a machine-learned concept, such as color, gender, or texture. There are also built-in analysis libraries like Audit-AI. Overall, there is a lot of work focused on identifying and measuring bias.

3. Method

To test for bias in our model, we will train two LSTM models with different pre-trained embedding layers on the same sentiment analysis task. The first model will implement GloVe embeddings while the second model will attempt to mitigate gender bias by using GN-GloVe.

GloVe embeddings use co-occurrence matrices to represent word meanings by telling us how often a particular pair of words occur together. Each value in a co-occurrence matrix is a count of a pair of words occurring together. GloVe then learns to encode the word probability ratio information in the form of word vectors. [3]

The Gender-Neutral variant of GloVe was created to mitigate gender bias inherited from pre-trained word embeddings by removing the "gender dimension"/characteristic from gender-neutral words.

With GN-GloVe, gender neutral words (i.e. scientist, nurse, doctor, programmer) will be represented by word vectors without the gender subspace, whereas gendered words have the ability to keep the gender dimension (i.e. King, Queen, daughter, son) – as its important to the word definition [4]

Therefore, when the trained model reaches the downstream task of sentiment analysis, it cannot base its predication off a shared gender trait that's semantically irrelevant [5]

4. Evaluation

We will use the Equity Evaluation Corpus to measure model bias across protected groups. This corpus is a curated set of text in where protected groups are added to templated sentences (ie: $< He >$ did., $< She >$ did., ...). [6] We then will measure whether the predicted sentiment for these sentence variants from our trained model is positive, neutral, or negative given the range [-1,1]. In addition, we will measure the intensity of the prediction with its valence measure. A valence measure closer to -1 indicates a negative sentiment prediction with high intensity, while a valence measure closer to 1 indicates a positive sentiment prediction with high intensity. We can then calculate bias based on differences in these sentiment measurements for same-template sentences.

5. Further Work

While our base project will focus on debiasing gender, we hope to address other protected groups like race or sexual orientation which are not as easily identifiable in unlabeled datasets. This would require either finding a labeled dataset or finding an algorithm that can apply such labels.

Also, we would like to explore building in debiasing techniques, such as the cability to augment dataset with gender-swapped names.

References

[1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning, Sep 2019.

[2] H. Gonen and Y. Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them, Sep 2019.

[3] S. R. Pennington, J. and C. Manning. Glove: Global vectors for word representation, Oct 2014.

[4] J. Zhao, Y. Zhou, Z. Li, W. Wang, and K.-W. Chang. Learning gender-neutral word embeddings, Aug 2018.

[5] G. A. T. S. H. Y. E. M. Z. J. M. D. B. E. C. K. Sun, T. and W. Wang. Mitigating gender bias in natural language processing: Literature review, Jul 2019.

[6] S. Kiritchenko and S. M. Mohammad. Examining gender and race bias in two hundred sentiment analysis systems, May 2018.