| Name(s)  **Nitya Parthasarathy** | Project Number **S0417** |
|---|---|

**Project Title**

## Evaluation of Gender Bias in Social Media Using Artificial Intelligence

**Abstract**

**Objectives/Goals**

The objective was to come up with a generic methodology to study bias in social media content. Specifically, the goal was to study the occurrence of gender stereotypes in large social media databases.

**Methods/Materials**

The research focuses on two publicly available datasets, the IMDB database of movie reviews (50,000 reviews) and the Amazon movie critique database (random sampling of 25,000 reviews)

The first part of this work comprises of a statistical and probabilistic study of stereotype occurrences employing relevant metrics from diverse areas. Next, two algorithms based on Artificial Intelligence (AI) techniques are introduced to further assess and validate the initial results.

**Results**

First, I studied the distance of gender words to gender stereotypes using statistical analysis. The results show how close the female gender words are to female gender stereotypes as compared to the male gender stereotypes. For the reviews under study, the statistical results demonstrate that female gender words exhibit distinct closeness/correlation to female stereotypes. Next, I developed machine intelligence algorithms using both Bayesian probabilistic formulae as well as Neural Networks to further check if the gender could explicitly be determined just by observing/analyzing stereotypes in words surrounding the gender word. The results in fact prove that a fairly accurate algorithmic prediction can be made on whether a sentence contains a male or female gender from the surrounding words.

**Conclusions/Discussion**

This research introduces novel metrics to study gender bias in any social media content. The conclusions are drawn based on a comprehensive approach with the use of multiple mathematical models, both statistical and algorithmic to provide an accurate and reliable evaluation. The results conclusively show the presence of female gender stereotypes. In fact, the female gender words in a sentence are shown to be accurately predictable. While this is also true for the male gender, the evidence there is not as conclusive as that for the female gender. Note that this method readily can be extended to study age and race related bias and hence is an important advancement in automatically studying all forms of bias or stereotypes in any general class of social media content/tweets/documents.

**Summary Statement**

My research has led to a set of algorithms which can be applied to detect all forms of bias in social media thereby being an important contribution to bias studies using Artificial Intelligence.

**Help Received**

I discussed results with Prof. Sameer Singh at UCI to gather new datasets to extend my work.