**White Paper Spring 2023**

**Name: Ryan Bui**

**Project Name: Jigsaw Unintended Bias in Toxicity Classification**

**1. Introduction**

This is a kaggle challenge created in 2019 that is a response to a previous kaggle challenge which also set out to detect toxicity. However, they noticed that many of the models from that challenge incorrectly learned to associate the names of frequently attacked identities as toxic. So, my goal is to create a model that accurately detects toxicity while also minimizing this unintended bias.

**2. Dataset:**

This dataset comes from kaggle and it is a collection of approximately 2 million comments from various new sites. This data set was collected by a plugin called “Civil Comments” which was a way for people to peer review comments posted on independent news websites. When they shut down in 2017 they released their archive of comments to the public. The comments in the data set have already been labeled. According to the Kaggle challenge page it was labeled by 10 people who each labeled the comments either: Not Toxic, Hard to Say, Toxic, or Very Toxic. The data provided is all csv’s and the files total in 2.38GB.

**3. Data Preparation:**

The first step in data mining on the Jigsaw dataset is to prepare the data. This involves cleaning and transforming the data to ensure that it is in a format that can be easily analyzed. Once the dataset has been downloaded, I will check for missing values, outliers, and inconsistencies. Any missing values should be imputed, and outliers should be removed or treated appropriately. I will also make sure my target column only has values that are 1’s and 0’s so that I can apply classification techniques properly. Since my input variable is essentially text I will have to preprocess it to make it easier for my model to train. This means lemmatizing the text, removing stop words, and removing punctuations. I will also have to tokenize the text as well.

**4. Exploratory Data Analysis**:

After the data has been prepared, the next step is to perform exploratory data analysis (EDA). EDA involves visualizing and summarizing the data to gain insights into the relationships between the variables. For the Jigsaw dataset, this may involve creating histograms of how many times certain identities are mentioned in each of the comments. I can also create Word Clouds to help visualize prominent words.

**5. Modeling**

The next step is to select a modeling technique. I will use three algorithms for the Jigsaw dataset, including **Naive Bayes Algorithm, logistic regression, and Neural Networks using transfer learning**. The goal of modeling here is to predict the type of iris flower based on the characteristics of the flower. The dataset includes comments that are rated Not Toxic, Hard to Say, Toxic, or Very Toxic and the target column represents a float number which is the ratio of how many people thought the toxic was toxic or worse. The modeling process involves splitting the data into training and testing sets, training the model on the training set, and then evaluating the model's performance on the testing set.

**6. Evaluation**

Model evaluation is an important step in data mining on the Jigsaw dataset. This involves calculating performance metrics, such as **accuracy, precision, recall, and F1 score**, to determine how well the model performs. The Kaggle data set also has their own set of metrics in which the competition will be graded on it composes of these: There will be a variety of metrics used to assess performance such as: Overall AUC: This is the ROC-AUC for the full evaluation set. BPSN (Background Positive, Subgroup Negative) AUC: Here, we restrict the test set to the non-toxic examples that mention the identity and the toxic examples that do not. BNSP (Background Negative, Subgroup Positive) AUC: Here, we restrict the test set to the toxic examples that mention the identity and the non-toxic examples that do not. To combine the per-identity Bias AUCs into one overall measure, we calculate their generalized mean as defined below:

