Evaluation of Gender Bias in Social Media Using Artificial Intelligence

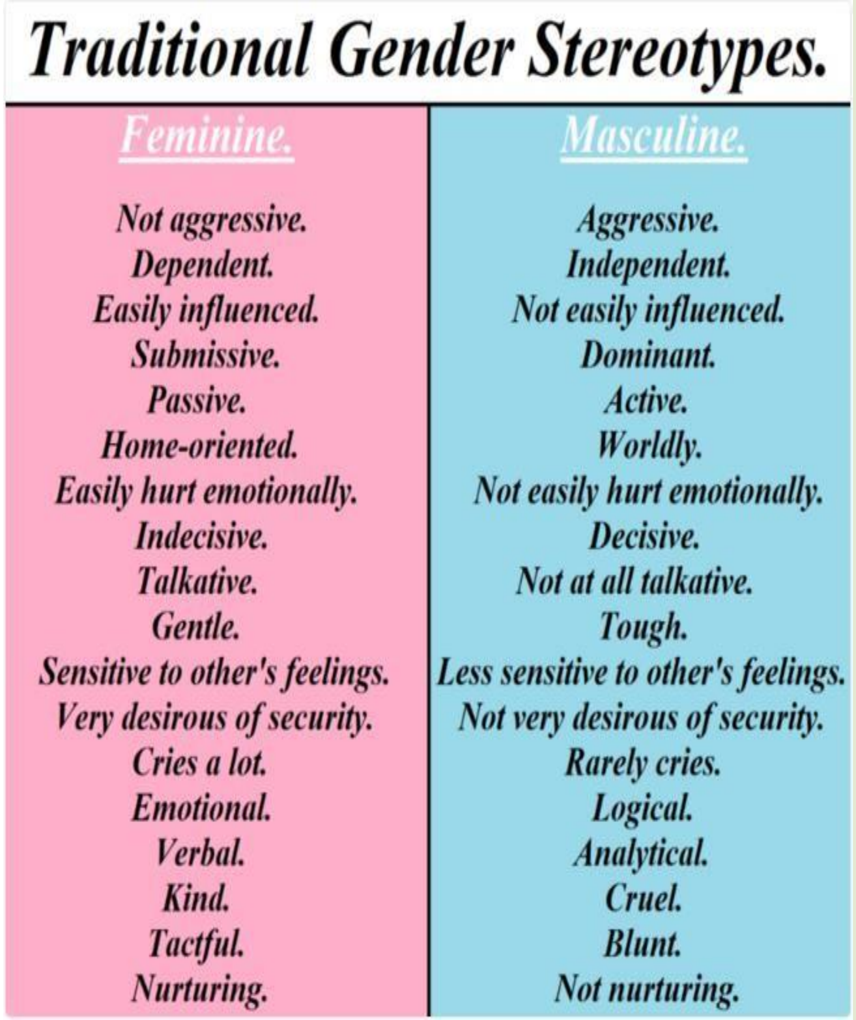
Nitya Parthasarathy

Presentation Overview

- Introduction
- Relevant Work
- Methodology
  - Statistical Analysis of Gender Bias
    - Statistical metrics to study gender word/stereotype frequency
  - AI based Algorithmic Analysis
    - Baye’s, Neural Networks and other AI based algorithmic approaches
- Results
- Conclusions and Future Directions

Introduction

A casual search on the internet brings up studies on various forms of bias



Traditional Gender Stereotypes.

Feminine.	Masculine.
Not aggressive.	Aggressive.
Dependent.	Independent.
Easily influenced.	Not easily influenced.
Submissive.	Dominant.
Passive.	Active.
Home-oriented.	Worldly.
Easily hurt emotionally.	Not easily hurt emotionally.
Indecisive.	Decisive.
Talkative.	Not at all talkative.
Gentle.	Tough.
Sensitive to other's feelings.	Less sensitive to other's feelings.
Very desirous of security.	Not very desirous of security.
Cries a lot.	Rarely cries.
Emotional.	Logical.
Verbal.	Analytical.
Kind.	Cruel.
Tactful.	Blunt.
Nurturing.	Not nurturing.

Example of some stereotypes listed in gender studies

Introduction: Problem Statement

- Are there implied societal and behavioral roles (either overt or subliminal) that are encouraged in social media?
  - If so, is this more prevalent for the male or female gender class?
  - How can one evaluate and quantify the degree of bias?
- Can one then develop Artificial Intelligence (AI) based and self-learning algorithms to identify and quantify any form of bias in social media?
  - Language Independence desired
  - "*BiasScore*" Fairness Metric
- Secondary question: how much of the bias does an AI program absorb while evaluating a target material and is this algorithm specific?

Relevant Work

- Related work is more observational with most performing statistical surveys
  - Word embeddings trained on Google news articles exhibit gender stereotypes
  - Wikipedia edits and sports journalism have also been shown to have bias in language
  - Gender inequality in movies/movie critiquing have been analyzed to evaluate the ratio of men to women
- In contrast, this work focusses on developing a comprehensive statistical as well as algorithmic framework
  - Sophisticated classifiers at a sentence level with applications to any social media.
  - Both syntactic and semantic constructions are leveraged to develop an unsupervised classifier to predict the gender of any mention from its context.
- Original in this regard
  - Big step forward in extending machine intelligence algorithms/advanced statistical metrics to new directions in behavioral and social sciences for analyzing biased language, text and interaction.

Methodology

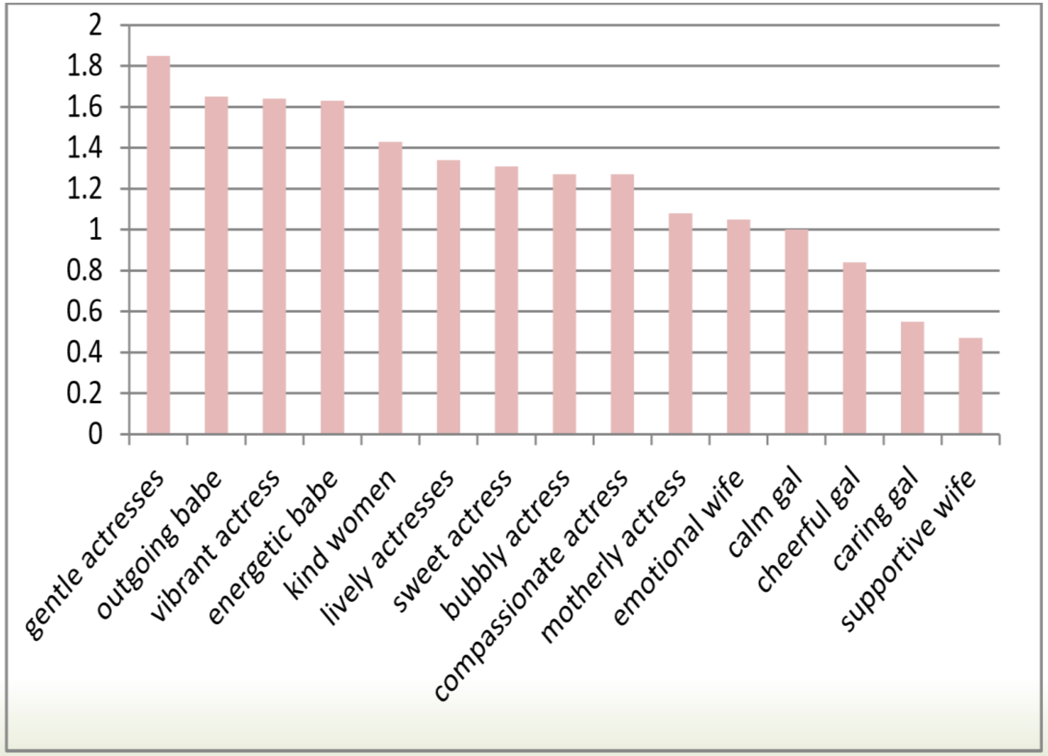
- Comprehensive statistical as well as algorithmic framework using sophisticated classifiers
- Statistical Inference using
  - Novel statistical metrics such as “Positivity” are introduced
  - Relate metrics such as “NMPI” from the field of Information theory
  - Metrics study the co-occurrence (proximity) of gender words and stereotypes for both female/male gender and examine their statistical distributions
- Algorithmic analysis using AI motivated by the game of Hide-and-Seek
  - Words around the gender word are used to create a model for the gender word/stereotype association
  - Existing and new classifiers developed to study
  - Gender word in sentence is deleted. Text is considered biased when the prediction of the gender word (using AI trained models and surrounding words) matches the actual gender
  - Uses the widely available and very large IMDB and Amazon movie reviews dataset

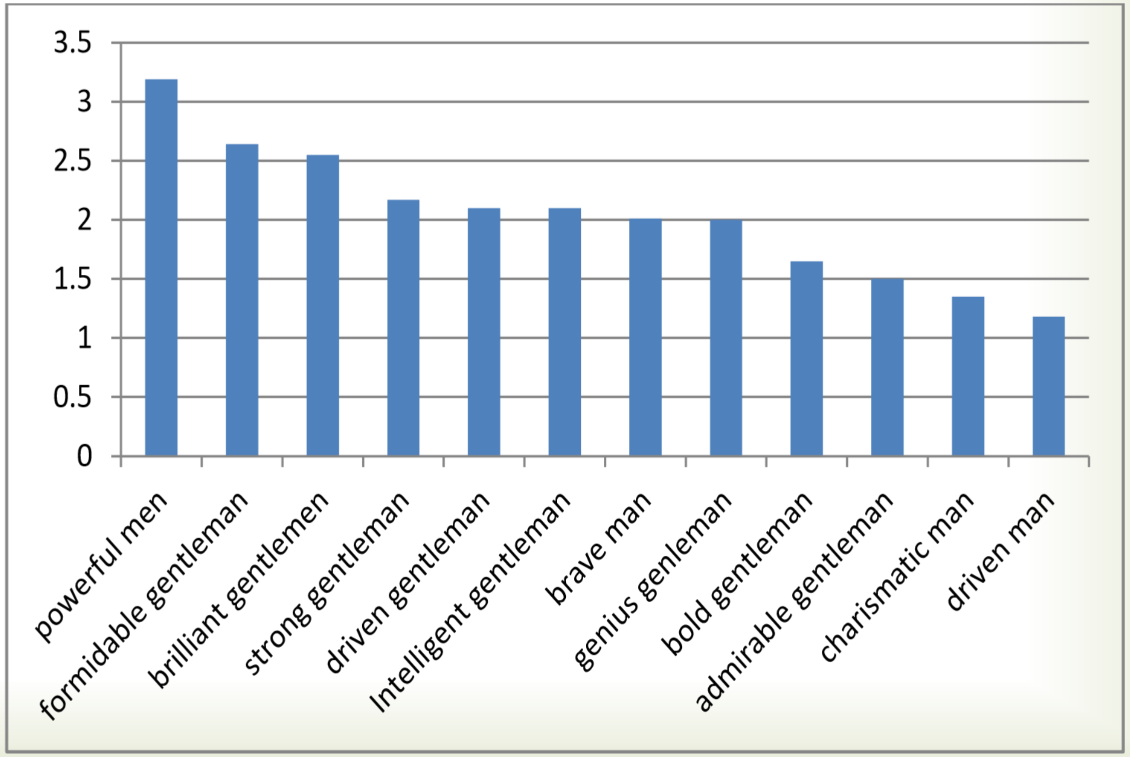
Evaluation Datasets

- Typically one of the most challenging parts of AI projects
  - Meaningful and relevant datasets
  - Large enough to achieve statistical confidence
- Publicly available movie review data sets used in this work
  - Widely available
- IMDB movie database
  - 25,000 reviews tagged with positive sentiment, 25,000 reviews tagged with negative sentiment and 50,000 unclassified reviews
  - From all genre and timespans
- Amazon movie database
  - ~8 million total unclassified reviews
  - Random sampling of 250,000 reviews used

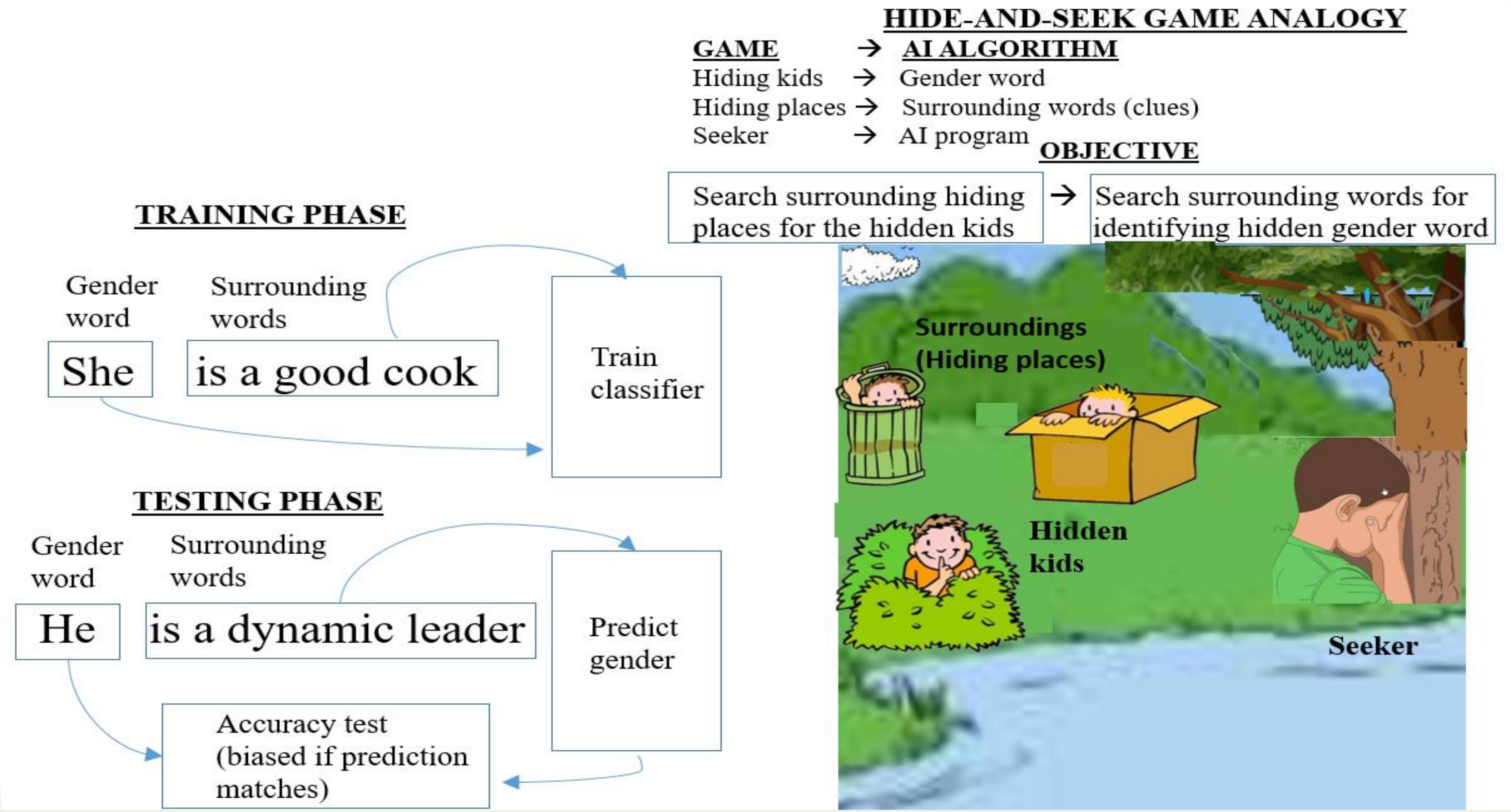
Statistical Analysis

- Consider the word “beautiful”. How frequently does this word occur in close conjunction (in terms of word distance) with a female description (a female gender indicator word such as “woman” or “lady”)?
- **PHRASE POSITIVITY** = *Probability of phrase occurrence in a document with overall positive sentiment – Probability of phrase occurrence in a document with overall negative sentiment*





AI technique inspired by Hide-and-Seek



Algorithmic Analysis using AI

- Bayesian classifiers are widely used as a feature learning algorithm in machine intelligence
- Suppose *Gw* = Gender word (set of *Mw* and *Fw*, the male and female gender word), ***Sw*** = Vector of words in the sentence surrounding the gender word, using Baye's formula …..

1. $\mathrm{P}(Gw/\boldsymbol{Sw}) = \mathrm{P}(\boldsymbol{Sw}/Gw) * \mathrm{P}(Gw)/\,\mathrm{P}(\boldsymbol{Sw})$
2. $(\prod_{j=1}^{n} \mathrm{P}(Sw_j/Mw) * \mathrm{P}(Mw)) > (\prod_{j=1}^{n} \mathrm{P}(Sw_j/Fw) * \mathrm{P}(Fw))$
3. $\sum_{j=1}^{n} \log(\mathrm{P}(Sw_j/Mw)) + \log(\mathrm{P}(Mw)) > \sum_{j=1}^{n} \log(\mathrm{P}(Sw_j/Fw)) + \log(\mathrm{P}(Fw))$

TRAINING

Step 1	**Loop through each sentence of every document …..**
Step 2	For each sentence, optionally delete "Stopwords" (words like "and", "if", "it" which carry no special information)
Step 2	Compute P(Sw_j/Mw) and P(Sw_j/Fw), where j = 1 … N Probability is computed with a word frequency estimate
Step 3	Compute P(Mw) and P(Fw) over the same set of training documents again with frequency estimates

TESTING

Step 1	**Loop through each sentence of a given document …..**
Step 2	For each sentence, optionally delete "Stopwords"
Step 3	Delete MaleWord (Mw) or FemaleWord (Fw) if it belongs to the pre-assigned male or female set of gender words
Step 4	Re-compute the gender word (whether it is Mw or Fw) based on Equation 3
Step 5	Count correct/wrong results to track statistics

AI based bias estimation: Some highlights ....

- NN model as in the paper by Mikalov called **Word2Vec** is extended for bias analysis
- Training/testing procedure similar to Baye's algorithm illustrated earlier
- An example from training: **Gentle + Woman** = **Naive !!**
  - Implies solution to the equation $\boldsymbol{Max_i\,(cos(w_i\,,\,Gentle)\;+\;cos(w_i\,,Woman))}$ is $\boldsymbol{w = Naive}$ where ***cos* = *cosine similarity*** and $\boldsymbol{w_i}$ refers to the all the words in the database

Gentle + Man = "Compassionate" is the closest word
Strong + Man = "Dignity"
Strong + Woman = "Naïve"
Providing + Man = "Successful"
Providing + Woman = "Delicate"
Caring – Man + Woman = "Emotional"

Word similarity highlights from Word2Vec Neural Net training

Other Classifiers

1) **Logistic Regression**: A special case of generalized linear regression where input, output relationship between the dependent/independent variables takes the form of a *"logit"* function
2) **Decision Tree**: A set of attributes are tested in the form of a multilevel tree. At each level of the tree, some function of the attribute is tested until a decision is arrived at
3) **Multi Layer Perceptron (MLP)**: A class of feedforward neural network with an input layer, output layer and at least 1 hidden layer. Connection weights are adapted through back propagation
4) **AdaBoost (Base Classifier: Decision Tree)**: An algorithm wherein the information gathered at every stage of the Adaboost algorithm is combined with the base classifier (decision tree)
5) **Ensemble Classifier (Hard Voting)**: Multiple learning algorithms (in this case, the above 4 classifiers along with Naïve Baye's) are combined with majority vote of the constituent classifiers are used to result in a decision
6) **Ensemble Classifier (Soft Voting)**: Same as above except that instead of hard voting, the average of the predicted probabilities ("soft vote") is provided as the class label with appropriate weighting
7) **A Bi-directional Long Short Term Memory (LSTM) Classifier:** A special class of recurrent Neural networks (RNN) which are capable of retaining long short term memory thereby extending the neural network ability

Metrics for bias evaluation

Precision = Total correct male gender / Total predicted male gender "*Precision"* metric for the male gender evaluation tells that among the predicted male gender words, how many were actually correct
Recall = Total correct male gender / Total male gender *Recall"* metric depicts the other side of predictions It shows that of the total male gender words, how many were correctly predicted.
F1 = 2*Precision*Recall / (Precision + Recall) "*F1 score"* merges "Precision" and "Recall" into a single metric using their harmonic mean.

Results (Note: Random guess has 50% accuracy)

Amazon Database	**Gender**	**Precision**	**Recall**	**F1**	**Accuracy**
Logistic Regression	Male	0.64	0.59	0.61	64.4%
	Female	0.65	0.69	0.67	
Decision Trees	Male	0.62	0.33	0.43	58.4%
	Female	0.57	0.81	0.67	
Ada Boost	Male	0.61	0.54	0.58	61.9%
	Female	0.62	0.69	0.66	
Naïve Bayes	Male	0.61	0.45	0.52	59.8%
	Female	0.59	0.73	0.66	
Multi-level Perceptron	Male	0.67	0.68	0.67	68.5%
	Female	0.70	0.69	0.70	

Results (Note: Random guess has 50% accuracy)

Amazon Database	**Gender**	**Precision**	**Recall**	**F1**	**Accuracy**
Ensemble (Hard Voting)	Male	0.65	0.52	0.58	
	Female	0.63	0.75	0.69	64.0%
Ensemble (Soft Voting)	Male	0.67	0.64	0.65	
	Female	0.68	0.72	0.70	67.8%
Word2Vec	Male	0.49	0.50	0.49	
	Female	0.56	0.54	0.55	52.5%
LSTM	Male	0.67	0.66	0.67	
	Female	0.63	0.65	0.64	65.4%

"BiasCheck": A web-based tool for automated evaluation of gender bias

1) Process the document to tag male or female gender words on a per sentence basis
2) Delete the gender word
3) Use the base classifier to estimate the gender word as well as gender word probability.
4) "*BiasScore*" is obtained by accumulating all the weighted individual sentence bias scores

$BiasScore* =$

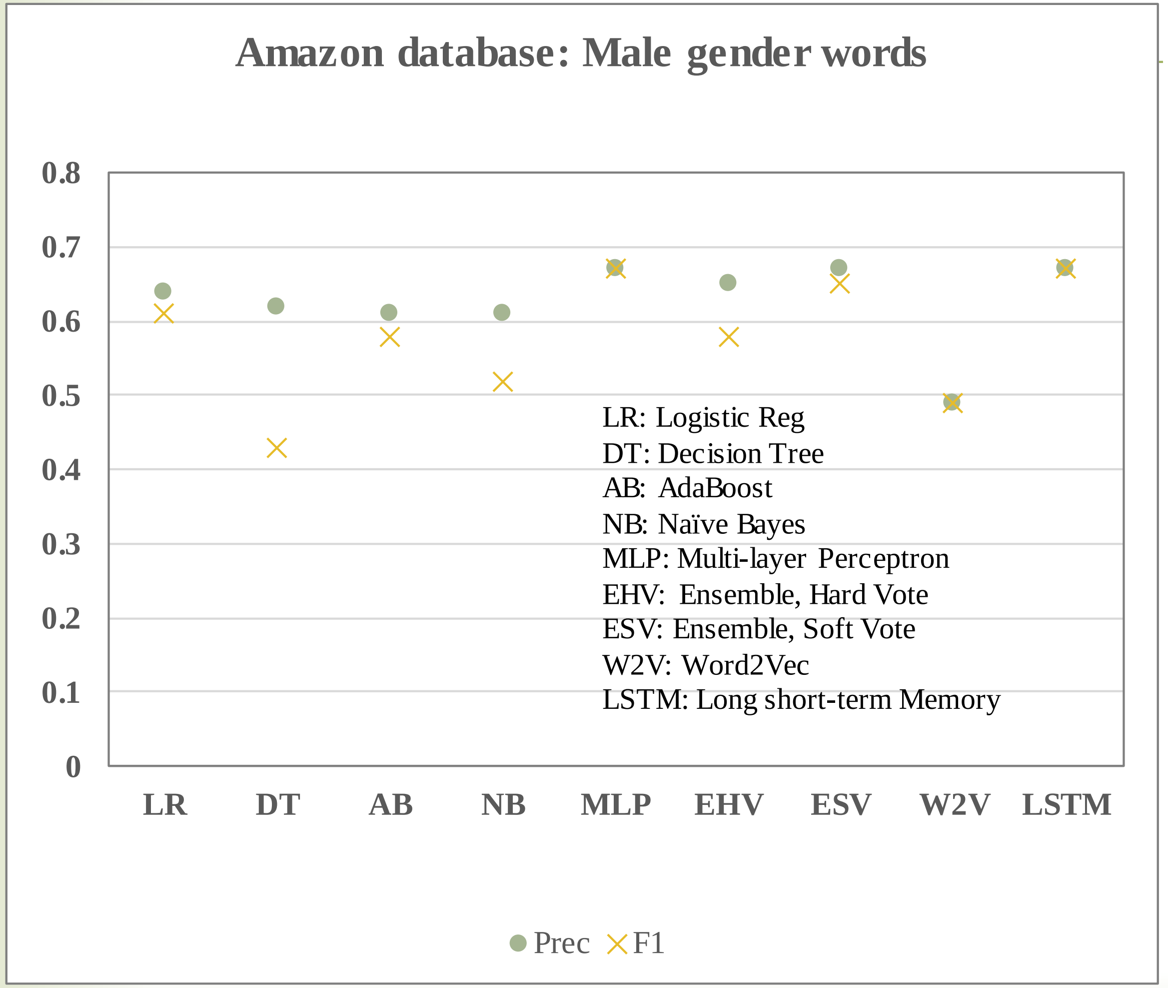
$\frac{1}{M}\sum_{j=1}^{M} Probability\ of\ j^{th}\ masked\ gender\ word\ prediction, M = total\ gender\ word\ instances.$

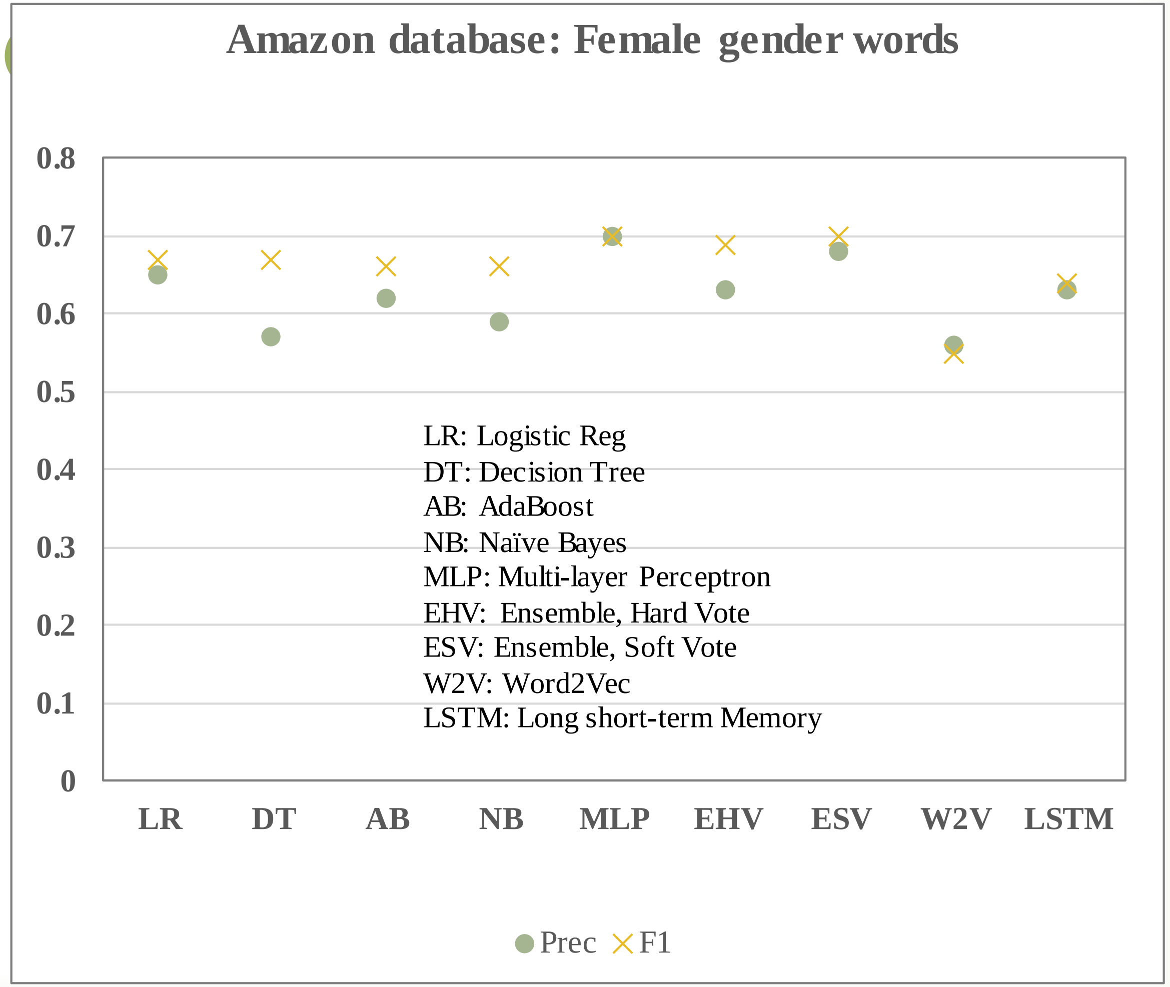
*Higher the confidence in predicting the gender, the higher the bias value (which ranges from 0 to 1)

NOTE 1: This technique can potentially further be refined to predict bias only on sentences that contain behavioral descriptors like adverbs and adjectives
NOTE 2: This technique extends to predict racial bias too

Classifier Results (Amazon)

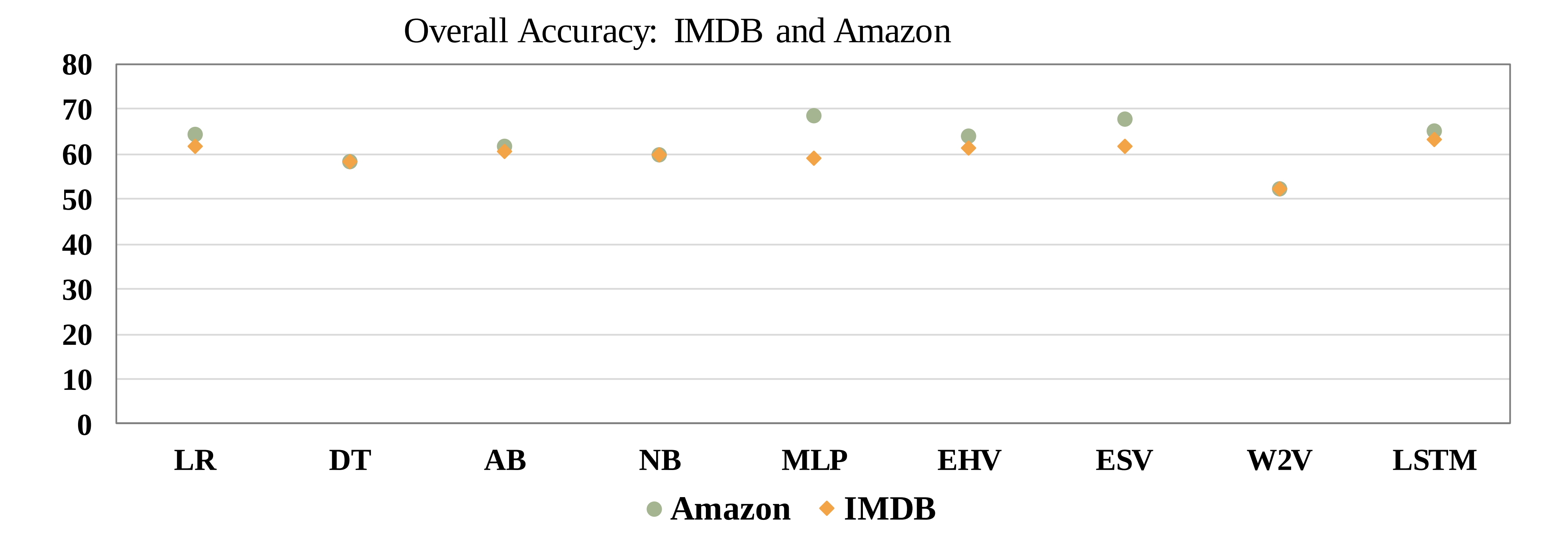
(Note: Random guess has 50% accuracy)





Overall Accuracy

(Note: Random guess has 50% accuracy)



<u>Conclusions</u>:

- Gender relatively easy to detect over multiple databases!
- All classifiers perform much better than random guess!
- Results correlate over multiple classifiers with small variations in detection accuracy
- All AI algorithms **<u>absorb</u>** the bias in the dataset. Use with caution!

Extensions to Other Bias/Stereotype Testing

Racial/Age/Social Strata bias evaluation:

Training:

- For every sentence in training set …
- Replace word denoting either race/age/social strata with a generic term "*Rw*"
- Train Nn on the training set to learn features for "*Rw*"

Testing:

- Delete words denoting race in a given sentence
- Delete gender words too if gender is additionally to be included
- Recompute if race/age/social strata word was present
  - NOTE: Can use techniques such as "Stemming" to improve detection accuracy
- Track correct/wrong evaluation statistics

"BiasCheck": A Web-Based Tool

- Process the document to tag male or female gender words on a per sentence basis
- Delete the gender word
- Use the base classifier to estimate the gender word as well as gender word probability.
- "*BiasScore*" is obtained by accumulating all the weighted individual sentence bias scores

BiasScore* =

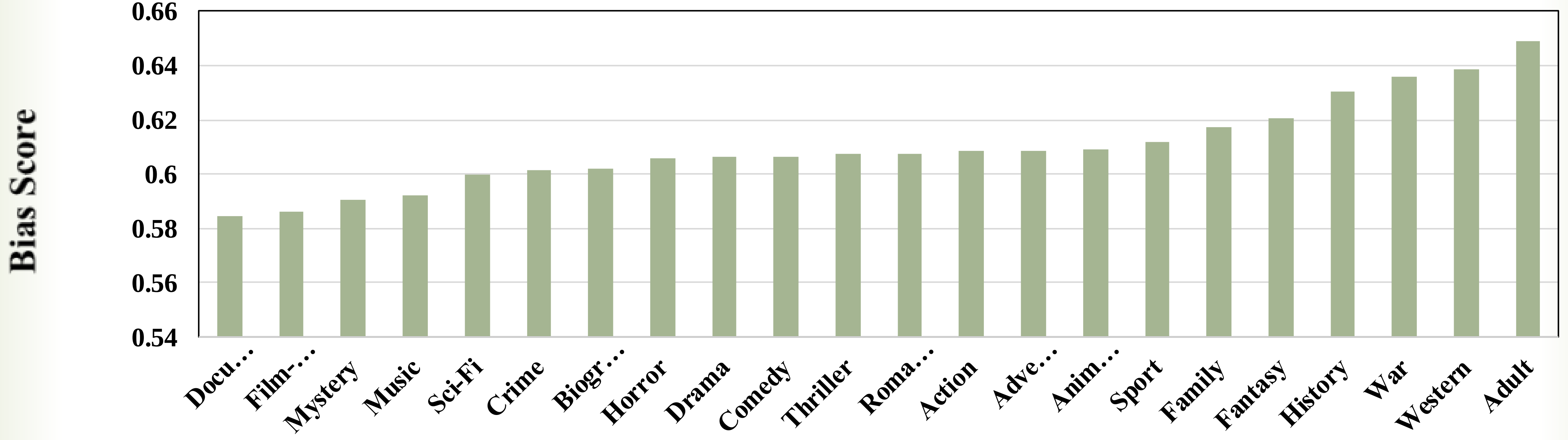
$$\frac{1}{M}\sum_{j=1}^{M} Prediction\ probability\ of\ jth\ masked\ gender\ word$$

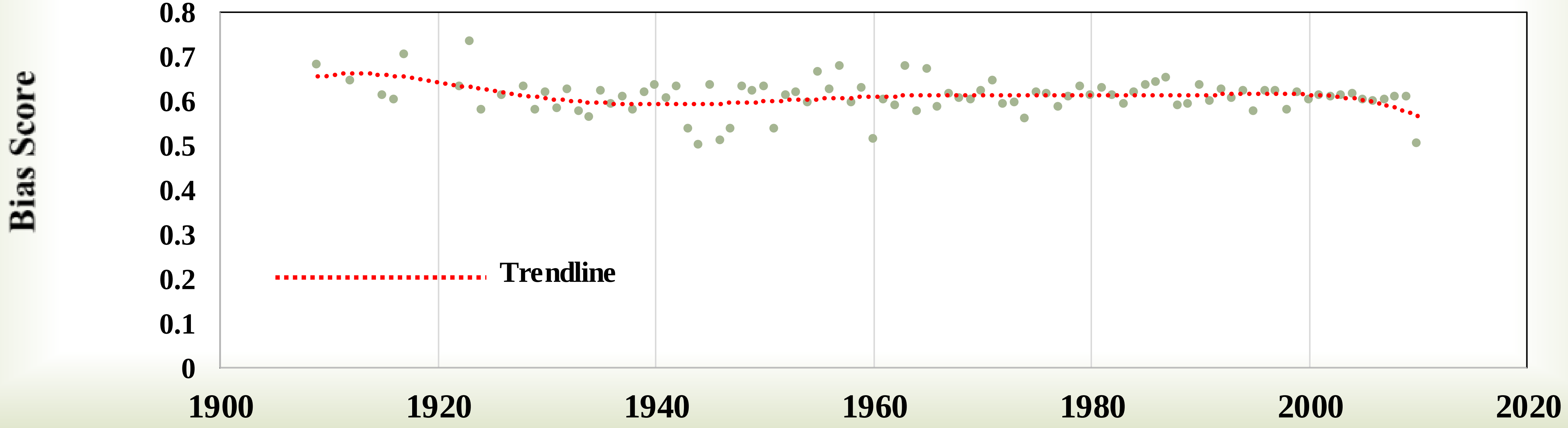
*Higher the confidence in predicting the gender, the higher the bias value (which ranges from 0 to 1)

Extensions to Sentiment Analysis:

- Sentiment analysis categorizes opinions in any text into positive, negative and neutral classes
- Initial results show less than 1% hit on analysis accuracy when gender bias words are deleted prior to evaluating sentiment
- Can further extend work to study impact of a specific bias/stereotype on sentiment analysis

Bias Score vs Genre and Year





Conclusions

- New statistical metrics, some of which were adapted from diverse areas such as "Information Theory" and "Language Modelling" were introduced to evaluate gender bias in social media
- Comprehensive results were provided to demonstrate the **<u>presence</u> of male and female gender stereotypes** in social media
- Female gender generally identified with "softer" roles while male gender was identified with "leadership" roles.
  - ❑ AI algorithms (using numerous classifiers) developed were able to pick up this bias
  - ❑ Interesting insights into social behavioral perceptions whereby a "providing man" was identified as "successful" whereas a "providing woman" was tagged as "delicate"!
- A new direction of applying statistical techniques and AI for "social good" has been established
  - ❑ Uncovered a rich set of topics for future study

Future Directions

- The techniques used in this work can further developed both on the *algorithmic* side as well as *socio-economic/behavioral* side.
- *Algorithmic* front: Interesting to see if it is possible to further improve the algorithm accuracy
    - What is the theoretical best that one could do?
    - Dynamic adaptation whereby the algorithms continue learning even on the evaluation data
    - Mobile app along the lines of "*SpellCheck*" which highlights/corrects the bias
    - Incorporate Part-of-speech (POS) models to potentially enhance the classifier accuracy
- *Socio-economic/behavioral* side: Are gender bias conforming movie commercially more successful?
- Extensions to racial, economic and political bias
    - Study correlation between biased speech and popularity

Acknowledgements

- I would like to gratefully acknowledge the invaluable discussions and guidance by Professor Sameer Singh in the department of Computer Science at the University of California, Irvine