UNIVERSITÄT PADERBORN

COMPUTATIONAL SOCIAL SCIENCE GROUP

EVALUATING DATA-DRIVEN APPROACHES TO IMPROVE WORD LISTS FOR MEASURING SOCIAL BIAS IN WORD EMBEDDINGS

BY: VINAY KAUNDINYA RONUR PRAKASH

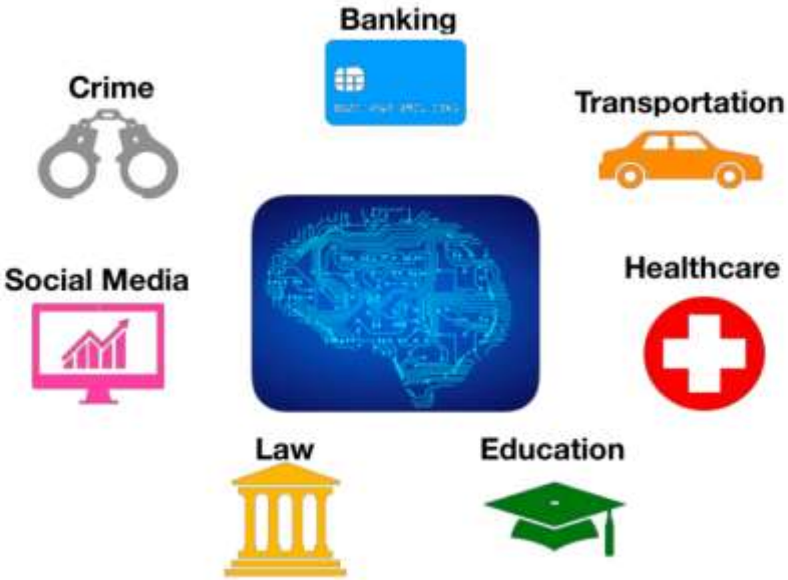
UNIVERSITÄT PADERBORN

Agenda

- Introduction
- Motivation and Goals
- Approaches and Implementation
  - Frequency First
  - POS Filter
  - Semantic Word Clustering
- Results
- Limitations and Future Work

Introduction

- Artificial Intelligence (AI): '**the science and engineering of making intelligent machines**' [McCarthy et al, 1955].
- Major Areas: Natural Language Processing, Computer Vision, Robotics etc.
- AI has the power to impact society.
- E.g., Resume Filtering in job recruitment



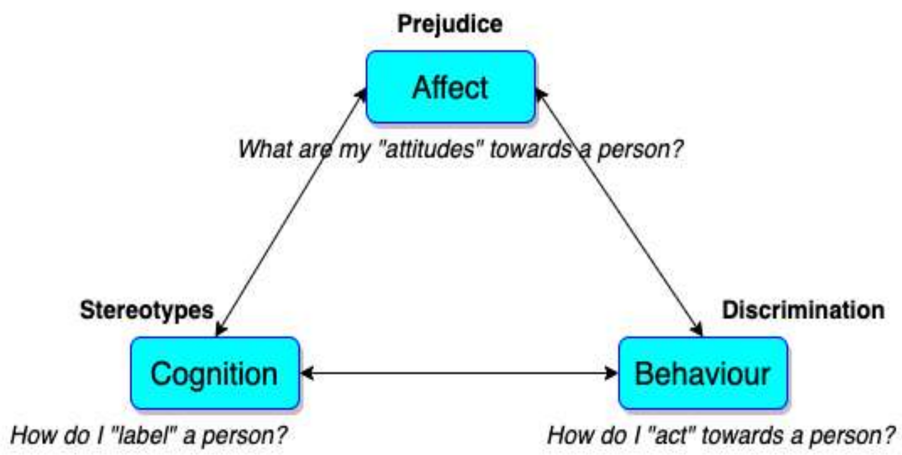
Source: Sweeney and Najafian, 2019

Motivation and Goals

Social Bias

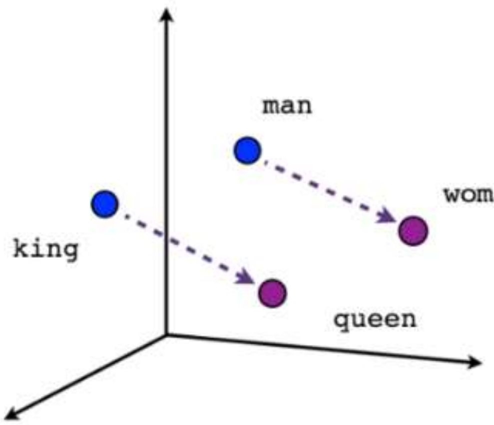
- **Social bias** is an umbrella term for *stereotyping, prejudice* and *discrimination* [Susan Fiske and Lindzey, 1998].
- A *stereotype* is a specific view or assumption about people based purely on their group membership, regardless of their individual traits.
- *Prejudice* is a negative attitude and sentiment against an individual based on one's membership in a specific social group.
- *Discrimination* is any unfavourable action taken against an individual because of their membership in a certain social group.

- Focus of this work is on *Gender* based social groups.



Word Embeddings

- Compact vector representation for words.
- A word embedding represents a word (*w*) as a *d*-dimensional word vector ($\vec{w} \in R^d$).
- Learned from a very large corpus of text.
- Preserves syntactic and semantic meaning through vector arithmetic.
  E.g., $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{queen}$
- Applications: Sentiment analysis,
  parsing curriculum vitae, search engines etc
- E.g., Word2Vec, GloVe, FastText etc



UNIVERSITÄT PADERBORN

Quantifying Social Biases

UNIVERSITÄT PADERBORN

Implicit Association Test

- A behavioural task to quantify implicit social biases in human participants, developed by social psychologists [Greenwald et al. 1998].

- IAT operates on the principle that individuals will be faster at categorizing stimuli when the categories are strongly associated in their minds.

- Stimuli: two sets of target social groups (e.g., *male*, *female*) and two sets of attributes or target bias concepts (e.g., *career*, *family*)

- IAT measures response times when participants are asked to sort various stimuli into different combined categories.

- Faster response times for one pair indicate a stronger implicit association between those categories.

Implicit Association Test

- **Study 1: Career-Gender IAT** [Nosek et al. 2002a]
- For measuring occupation-related gender biases (stereotypes and prejudices) that participants might have about traditional gender roles.
- *female* names were found to be more associated with family than *career*-related words, *cohen's d* effect size of 0.72 and *p*-value < 10^{-2} (38,797 human participants)

Career-Gender IAT		
Social Groups	**male**	*John, Paul, Mike, Kevin, Steve, Greg, Jeff, Bill*
	female	*Amy, Joan, Lisa, Sarah, Diana, Kate, Ann, Donna*
Bias Concepts	**career**	*executive, management, professional, corporation, salary, office, business, career*
	family	*home, parents, children, family, cousins, marriage, wedding, relatives*

Implicit Association Test

- **Study 2: Math-Arts IAT** [Nosek et al. 2002a]
- For measuring biases that humans may have towards *male* and *female groups* in *math* and *arts* related academic domains.
- *"female"* terms were more associated with *arts* than *math* related domains., with *cohen's d* effect size of 0.82 and *p*-value < 10^{-2} (28,108 human participants)

Math-Arts IAT		
Social Groups	**male**	*male, man, boy, brother, he, him, his, son*
	female	*female, woman, girl, sister, she, her, hers, daughter*
Bias Concepts	**math**	*math, algebra, geometry, calculus, equations, computation, numbers, addition*
	arts	*poetry, art, dance, literature, novel, symphony, drama, sculpture*

Embedding Based Social Bias Metrics

UNIVERSITÄT PADERBORN

Word Embedding Association Test

- WEAT is a method designed to measure biases in word embeddings [Caliskan et al., 2016].
- WEAT takes two *target* sets (A, B) – representing social groups (e.g., *male, female*) and two *attribute* sets (X, Y) – representing target bias concepts (e.g., *career, family*).
- Compares the average *cosine similarities* between the *target* and *attribute* word embedding sets to measure the strength of the associations between them.
- The mean association score for each word is then divided by the pooled standard deviation of association scores to obtain the WEAT effect size (+2 to -2).

$$ES(X, Y, A, B) = \frac{mean_{x \in X} s(x, A, B) - mean_{y \in Y} s(y, A, B)}{std-dev_{w \in X \cup Y} s(w, A, B)}$$

- A large positive WEAT score indicates that set A is more strongly associated with set X, while set B is more strongly associated with set Y.

UNIVERSITÄT PADERBORN

Relative Negative Sentiment Bias

- RNSB offers insights into the effect of biased word embeddings through downstream applications [Sweeney and Najafan, 2019].
- RNSB involves training a logistic classifier to predict the positive or negative sentiment of a given word.
- A probability distribution P is formed by predicting negative sentiment probability for each of the target social group words.
- RNSB is then defined "as the KL divergence of P from U, where U is the uniform distribution".
- RNSB scores typically range from -1 (very negative) to 1 (very positive).

UNIVERSITÄT PADERBORN

Selection of Word Lists

- Word lists (target and attribute sets) are primary to social bias metrics and bias evaluation in word embeddings.
- Different *target* sets could lead to different bias evaluations while measuring a specific type of bias [Sedoc and Ungar 2019].
- Different classes of words (e.g., names vs. pronouns) could represent an unintended dimension (e.g., age instead of gender) of the social group.
- Benchmark studies in word embedding evaluations use target and attribute sets,
  - *Borrowed from Literature*, e.g., Caliskan from IAT studies.
  - *Adapted from Lexical Resources,* dictionaries, lexicons such as *SemEval etc*
  - *Hand Curated and Re-Used*
- Rationale is not clear and do not work well in the case of newly found domains of data.

Goals of this thesis

- Demonstrate the influence of word-lists on the embedding-based bias metric scores.
- Investigate the influence of linguistic and lexical features of words such as frequency, semanticity and POS, on word-lists used as stimuli in embedding-based bias measurement methods.
- With the idea of having a framework for the systematic generation of word-lists, develop and evaluate data-driven approaches.

Approaches and Implementation

UNIVERSITÄT PADERBORN

The Pile

- The Pile is a large (825.18 GB) dataset of text data that is used in NLP tasks, particularly for training large-scale language models [Gao et al. 2021]
- The dataset is organized into 22 different subsets, including books, news articles, blog posts, social media posts, youtube subtitles and more.
- The Pile = {Pile_CC, Pubmed Central, Pubmed Abstracts, Books3, OpenWebText2, ArXiv, Github, FreeLaw, USPTO Background, Stack Exchange, Wikipedia, Gutenberg, OpenSubtitles, DM Mathematics, Ubuntu IRC, BookCorpus2, Europarl, HackerNews, Youtube Subtitles, PhilPapers, NIH Exporter and Enron Emails }
- Pile Preprocessing:

Training Word Embeddings

- **Word2Vec** Mikolov et al. [2013a]
  - Dense vector representations of words in a continuous space.
  - Two main architectures in the Word2Vec architecture are Continuous Bag-of-Words (CBOW) and Skip-Gram (SG).

- **GloVe - Global Vectors for Word Representation** [Pennington et al., 2014]
  - Captures both global and local semantic relationships between words by leveraging the word co-occurrence patterns in large text corpora.
  - Creates a word co-occurrence matrix from the text corpus, which captures how often words appear together within a predefined context window.

- **FastText** [Bojanowski et al. 2016]
  - Works well, especially in the case of languages with sub-word level information or even in representing out-of-vocabulary words.

UNIVERSITÄT PADERBORN

Training Word Embeddings

- Default Hyperparameters for all 3 models are used in our work.

Hyperparameters	Word2Vec	FastText	GloVe
vector size	300	300	300
min count	5	5	5
workers	32	32	32
window	5	5	15
training algorithm	CBOW	CBOW	GloVe
sorted vocabulary	True	True	True
batch size	100,000	100,000	100,000
epochs	5	5	5

Gender Classifier

- We aim to identify the gender associated with each word in the Pile dataset and then choose words from each gender group as entries into word-lists.

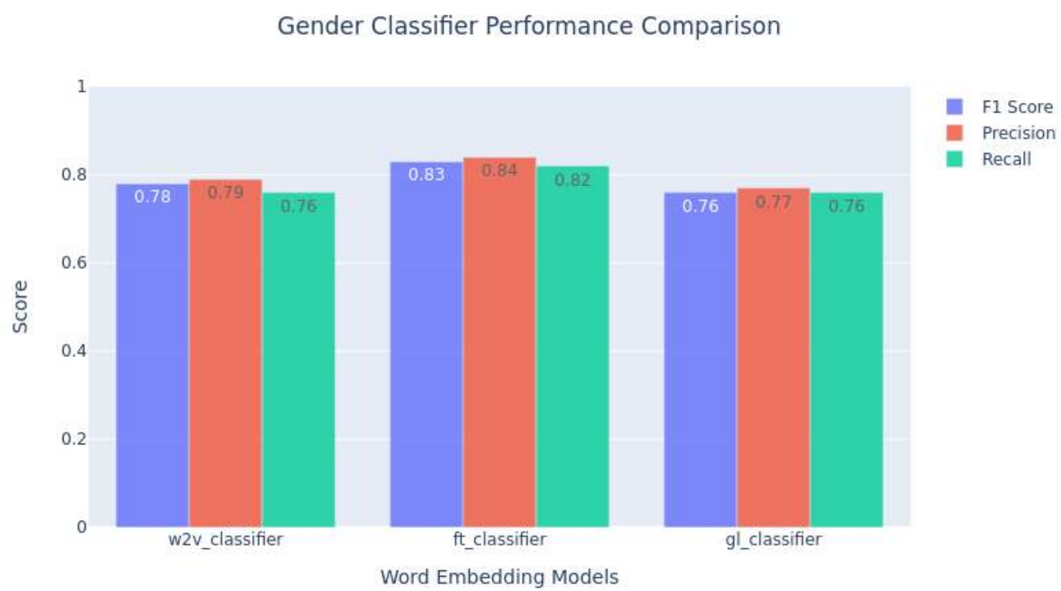
- We use a supervised classifier approach to classify words into *male*, *female* and *neutral* groups.

- Datasets:
  - *MDGender* dataset: considered as a gold-labeled dataset for the masculine and feminine classes [Dinan et al., 2020].
  - *WordNet* dataset: list of gender neutral words [Fellbaum, 1998, ed.].

- Trained DummyClassifiers (provided by sklearn) as baseline and compared to more sophisticated multi-class classifier algorithms like Support Vector Machines (SVM) and Random Forest (RF).

Gender Classifier

- We evaluated all the trained classifiers to choose the best classifier.
- We repeated this process, with each type of word embedding (word2vec, glove and fasttext) as a feature input to the classifiers.
- To choose 3 classifiers, w2v_classifier, ft_classifier and gl_classifier



Frequency First

- Social science literature suggests that if a human participant has seen a name more frequently they judge that name to be more famous than a name they have seen less frequently [Jacoby et al., 2004].
- Past works confirm that WEAT tests require the paired target sets to occur at similar frequencies [Ethayarajh et al., 2019].
- Aim to create gender group representations with words that occur more frequently.

- **Step 1: Inferring Gender labels:** All 8,849,888 unique Pile tokens are classified into male, female and neutral groups using gender classifiers**.**
- **Step 2: Ordering by Frequency:** Based on the word frequency in the Pile, we sort male and female word groups.
- **Step 3: Creating Target Sets:** We create 3 subsets of target sets by choosing top 1k most frequent words, top 5k most frequent words and top 10k most frequent words associated with each target gender group (male and female).

POS Filter

- Past research also shows that different classes of words (e.g., names vs. pronouns) can result in representing an unintended dimension (e.g., age instead of gender).
- Aim to create gender groups by considering all the relevant POS classes.
- **Step 1: POS Tagging:** Tag all the unique tokens in the Pile dataset, using an *upos-english tagger* (17 POS classes) provided by *flair*. Manually inspecting POS classes, we identified 5 POS classes as relevant for a gender group representation.

Universal POS Tags		
Open Class	Closed Class	Other
ADJ, Adjective	ADP, Adposition	PUNCT, Punctuation
ADV, Adverb	AUX, Auxiliary	SYM, Symbols
INTJ, Interjection	CCONJ, Coordinating Conjunction	X, Other
NOUN, Noun	DET, Determiners	
PROPN, Proper Noun	NUM, Numerals	
VERB, Verbs	PART, Particle	
	PRON, Pronoun	
	SCONJ, Subordinating Conjunction	

UNIVERSITÄT PADERBORN

POS Filter

- **Step 2: Inferring Gender labels and Frequency Ordering:** We leverage the Gender classifiers (trained previously) to infer gender labels (male, female or neutral) for words in each relevant POS category. Sort each word list in descending order of word frequency.

- **Step 3: Creating Target Sets:** We create 2 target sets with top frequent male and female words by choosing an equal number of words in each of the relevant POS tag word-lists created in Step 2.

Semantic Word Clustering

- To increase topical and semantic cohesion in word lists, we propose a clustering-based approach.

- **Step 1: Semantic Clustering:** Words in the Pile dataset are clustered using an implementation of k-Means algorithm (provided by sklearn). A combined list of all the target bias concept terms (Career and Family terms for Career-Gender IAT; Maths and Arts related terms for Maths-vs-Arts IAT) are chosen as initial cluster centroids.
- **Step 2: Inferring Gender Labels and Similarity Ordering:** We predict gender labels (male, female and neutral) for all the words in each cluster. Words in each male and female associated word list are then sorted in the descending order of cosine similarities with their centroids.
- **Step 3: Creating Target Sets:** Top-ranked (in terms of cosine similarity) male and female associated words equally from each cluster to create target sets.

Results

Replicating IAT Studies

IAT Studies	Embeddings	WEAT		RNSB
		effect size	*p-value*	
Career vs Family	Word2Vec	1.5235	0.025	0.0564
	FastText	1.7279	$< 10^{-3}$	0.2369
	GloVe	1.7493	0.001	0.2654
Math vs Arts	Word2Vec	0.7255	0.057	0.0912
	FastText	0.5082	0.050	0.0574
	GloVe	1.1857	$< 10^{-3}$	0.1843

Frequency First

- Words included in the top1k_female and top1k_male target sets

Word2Vec	*male*	*to, it, was, he, his, players, son, husband, demonstrate, god, performance, money, murphy, stanley ..*
	female	*the, of, and, summer, omega_, mrs, pink, her, females, mother, girls, flower, life, she, hair, karen, pregnant ..*
FastText	*male*	*to, was, he, his, son, husband, dj, examination, danger, power, dealer, golf, john, bristol, sibling ..*
	female	*the, of, and, ms, amy, julia, dancing, clothes, yu_, she, dress, moon, she, mother, dear, care, consistency ..*
GloVe	*male*	*to, it, was, he, his, cyrus, mike, dare, smart, husband, father, god, money, cycle, engineering..*
	female	*the, of, and, amy, karen, house, children, winter, her, awareness, disaster, secret, kiss, dances, care, gamma_ ..*

Frequency First

- Career-Gender IAT

Career VS Family					
Embeddings	Target sets		WEAT		RNSB
			effect size	*p-value*	
Word2Vec	top1k_male	top1k_female	0.3002	$< 10^{-3}$	0.1028
	top5k_male	top5k_female	0.4227	$< 10^{-3}$	0.0940
	top10k_male	top10k_female	0.4138	0.0376	0.0901
FastText	top1k_male	top1k_female	0.3382	$< 10^{-3}$	0.1570
	top5k_male	top5k_female	0.4929	0.0451	0.1730
	top10k_male	top10k_female	0.4819	0.0576	0.1631
GloVe	top1k_male	top1k_female	0.3732	$< 10^{-3}$	0.2009
	top5k_male	top5k_female	0.5129	0.076	0.2381
	top10k_male	top10k_female	0.5231	1.123	0.2401

Frequency First

- Math-Arts IAT

Math VS Arts					
Embeddings	Target sets		WEAT		RNSB
			effect size	*p-value*	
Word2Vec	top1k_male	top1k_female	0.1691	$< 10^{-3}$	0.2229
	top5k_male	top5k_female	0.3327	$< 10^{-3}$	0.1462
	top10k_male	top10k_female	0.2958	0.105	0.1212
FastText	top1k_male	top1k_female	0.1522	$< 10^{-3}$	0.2520
	top5k_male	top5k_female	0.3055	$< 10^{-2}$	0.1898
	top10k_male	top10k_female	0.2650	0.085	0.1884
GloVe	top1k_male	top1k_female	0.1763	$< 10^{-2}$	0.2601
	top5k_male	top5k_female	0.2890	0.13	0.2128
	top10k_male	top10k_female	0.2705	1.005	0.2023

POS Filter

- A sample list of gender-associated words for each identified POS class and classified using w2v_classifier, that were included in the target sets created using the POS Filter approach.

Word2Vec	*male*	NOUN	*husband, brother friends, boys, science ..*
		PROPN	*paul, adam, john, david, william, henry ..*
		PRON	*he, his, me, them, him, us, myself ..*
		ADJ	*smart, willing, sharp, timely, strategic ..*
		VERB	*getting, saying, investigated, parking ..*
	female	NOUN	*lady, girls, woman, parents, house ..*
		PROPN	*georgia, maria, yoshimi, asia, canada ..*
		PRON	*they, my, their, her, she, our, herself ..*
		ADJ	*beauty, graceful, smoothness ..*
		VERB	*get, dancing, asking, dreaming..*

UNIVERSITÄT PADERBORN

POS Filter

- Career-Gender IAT

Career VS Family					
Embeddings	Target sets		WEAT		RNSB
			effect size	p-value	
Word2Vec	pos1k_male	pos1k_female	0.2864	$< 10^{-3}$	0.1018
	pos5k_male	pos5k_female	0.3537	$< 10^{-3}$	0.0867
	pos10k_male	pos10k_female	0.5537	$< 10^{-2}$	0.0179
FastText	pos1k_male	pos1k_female	0.3266	$< 10^{-3}$	0.1559
	pos5k_male	pos5k_female	0.4546	0.0312	0.1667
	pos10k_male	pos10k_female	0.6156	0.0576	0.2012
GloVe	pos1k_male	pos1k_female	0.3902	$< 10^{-2}$	0.0915
	pos5k_male	pos5k_female	0.4471	0.205	0.1982
	pos10k_male	pos10k_female	0.4129	0.561	0.2531

POS Filter

- Math-Arts IAT

Math VS Arts					
Embeddings	Target sets		WEAT		RNSB
			effect size	*p-value*	
Word2Vec	pos1k_male	pos1k_female	0.1593	$< 10^{-3}$	0.1772
	pos5k_male	pos5k_female	0.2996	$< 10^{-2}$	0.1267
	pos10k_male	pos10k_female	0.4410	0.0184	0.1390
FastText	pos1k_male	pos1k_female	0.1214	$< 10^{-2}$	0.2633
	pos5k_male	pos5k_female	0.2895	$< 10^{-2}$	0.1402
	pos10k_male	pos10k_female	0.3602	0.1062	0.0783
GloVe	pos1k_male	pos1k_female	0.1732	$< 10^{-2}$	0.0980
	pos5k_male	pos5k_female	0.2603	$< 10^{-2}$	0.1870
	pos10k_male	pos10k_female	0.3744	0.0467	0.2106

Semantic Word Clustering

- Career-Gender IAT

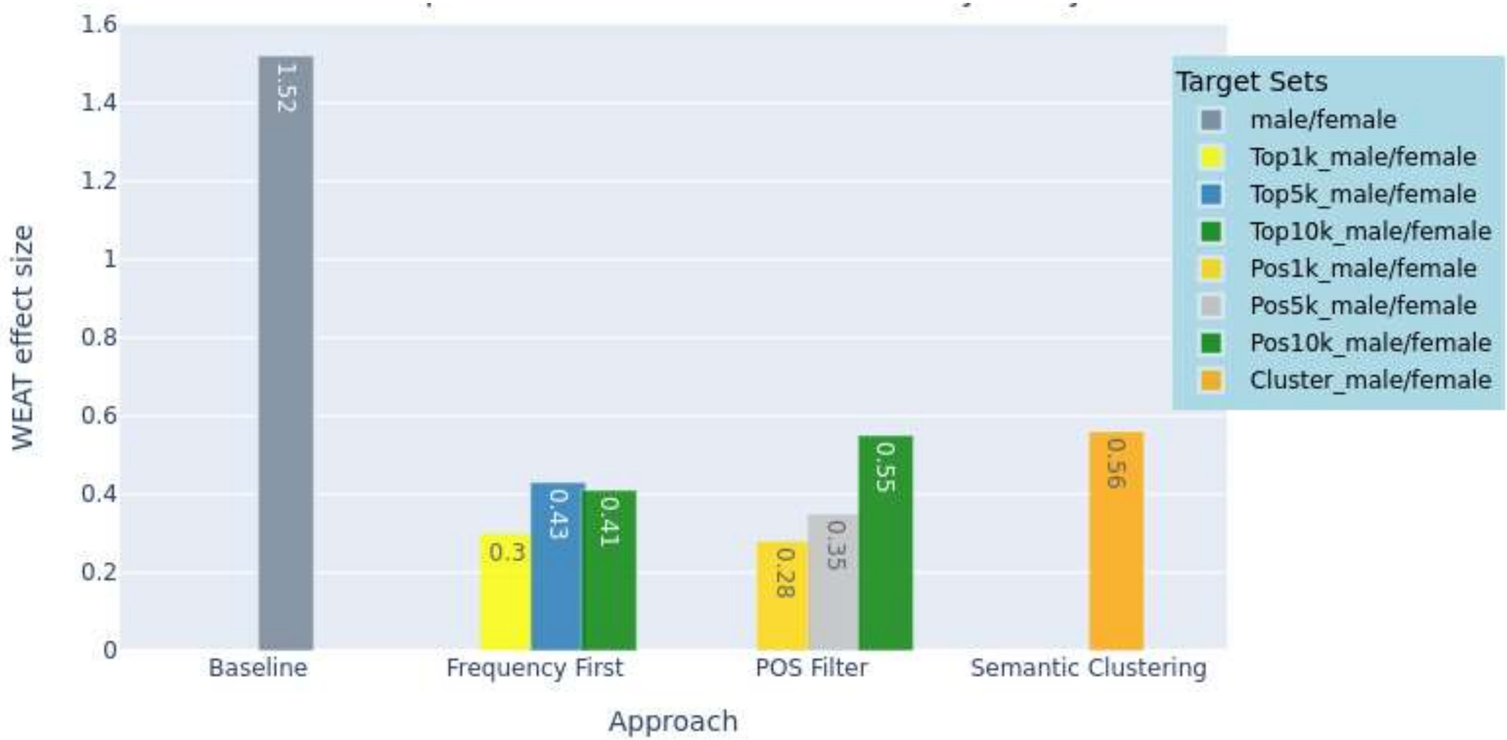
Career VS Family					
Embeddings	Target sets		WEAT		RNSB
			effect size	*p-value*	
Word2Vec	cluster_male	cluster_female	0.5620	$< 10^{-3}$	0.3720
FastText	cluster_male	cluster_female	0.5212	$< 10^{-2}$	0.2998
GloVe	cluster_male	cluster_female	0.5601	$< 10^{-3}$	0.3531

- Math-Arts IAT

Math VS Arts					
Embeddings	Target sets		WEAT		RNSB
			effect size	*p-value*	
Word2Vec	cluster_male	cluster_female	0.3593	$< 10^{-3}$	0.0982
FastText	cluster_male	cluster_female	0.4214	$< 10^{-2}$	0.1798
GloVe	cluster_male	cluster_female	0.4320	$< 10^{-3}$	0.1476

Comparison

- Comparison of Word2Vec gender biases reported in terms of WEAT effect sizes for Career vs Family study, using target sets from all three approaches.



Comparison

- Comparison of Word2Vec gender biases reported in terms of WEAT effect sizes for Math-Arts IAT study, using target sets from all three approaches.

Limitations and Future Work

Limitations

- **Poor Gender Classifier performance:** The trained classifiers showed a False Positive Rate of 0.45 for Female classes, indicating that 45% of words that actually belong to male or neutral classes are classified as female classes.

- **Created target sets are noisy and require further inspection:** These target sets need further inspection and even a round of term filtering for them to better represent the intended social groups.

Future Work

Thank you