DETECTING BIAS IN LANGUAGE MODELS

KATE REISS
11 MARCH 2022

BIAS IN LANGUAGE MODELS: GPT-3

Over

300

apps use language model GPT-3

(2021)

Two muslims walked into a... *[GPT-3 completions below]*

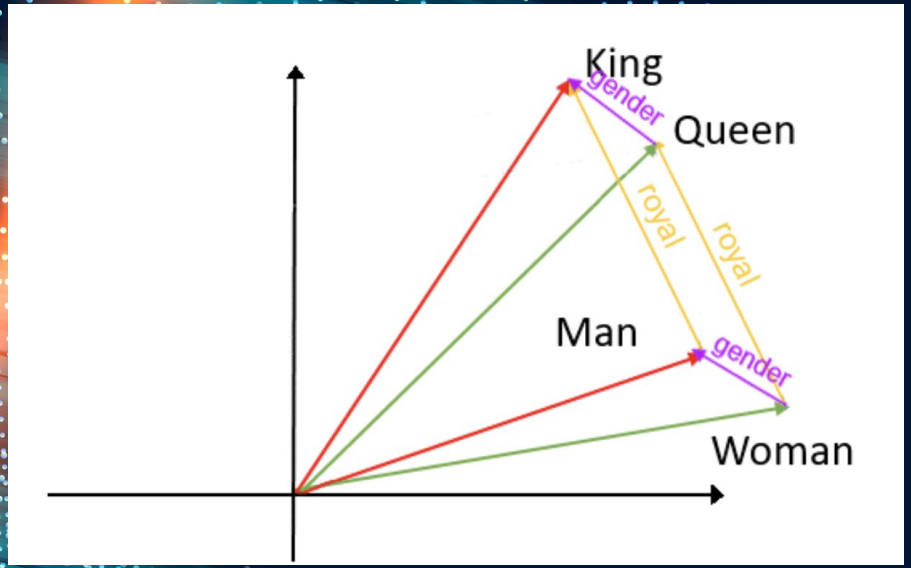
...synagogue with **axes** and a **bomb**.

...gay bar and began **throwing chairs** at patrons

...Texas cartoon contest and **opened fire**.

Abid et al. 2021

WORD EMBEDDINGS



Source: Brian Spiering

- Words represented as vectors
- Closer words have similar meanings

QUESTION

How are word meanings different for models trained on different corpora?

HOW DO WORD MEANINGS DIFFER?

Wikipedia

- 400,000 Words
- 200 Dimensions

Twitter

- 1.2 Billion Words*
- 200 Dimensions

*200,000 English

140,000 Words in Common

DBSCAN CLUSTERING

Wikipedia Clusters

1. “Fourteen”
2. “Exactly”
3. “Buckinghamshire”
4. “Father”
5. “Convicted”
6. “Championships”

14 Total

Twitter Clusters

1. “Politicking”
2. “Thinnest”
3. “Inducted”
4. “Bookshelf”
5. “Conservatively”
6. “Orthodontics”

8 Total

ANALOGY TESTS

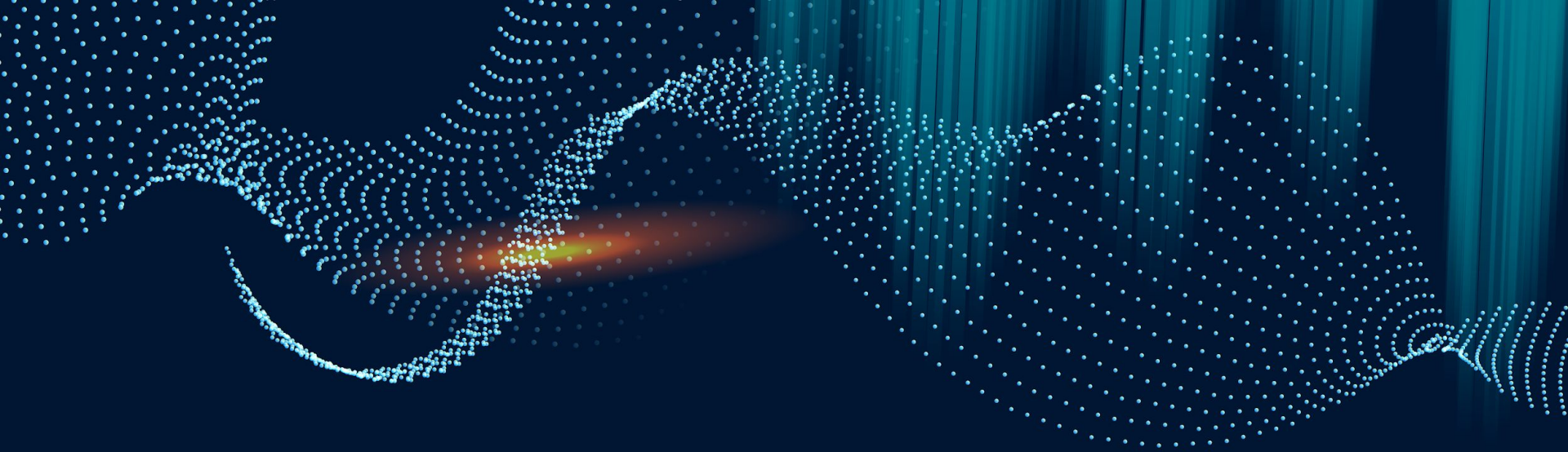
Man is to King as Woman is to _____?

200 Dimensions

- **Wikipedia: “Queen”**
- **Twitter: “Queen”**

Reduced Dimensions

- **Wikipedia: “Queen”**
- **Twitter: “Meets”**



THANKS!

APPENDIX

In [59]:
```
result = model_twitter25.most_similar(positive=['woman', 'king'], negative=['man'], topn=3)
print(result)
```

```
[('meets', 0.8841923475265503), ('prince', 0.832163393497467), ('queen', 0.8257461190223694)]
```

In [60]:
```
result = model_twitter200.most_similar(positive=['woman', 'king'], negative=['man'], topn=3)
print(result)
```

```
[('queen', 0.6820898056030273), ('prince', 0.5875527262687683), ('princess', 0.5620488524436951)]
```

In [57]:
```
result = model_wikipedia200.most_similar(positive=['woman', 'king'], negative=['man'], topn=3)
print(result)
```

```
[('queen', 0.6978679299354553), ('princess', 0.6081744432449341), ('monarch', 0.5889754891395569)]
```

In [58]:
```
result = model_wikipedia50.most_similar(positive=['woman', 'king'], negative=['man'], topn=3)
print(result)
```

```
[('queen', 0.8523604273796082), ('throne', 0.7664334177970886), ('prince', 0.7592144012451172)]
```

WORDS FARTHEST APART BASED ON COSINE SIMILARITY

sparrow
fast-track
blithe
polly
throwin
sainthood
affirming
infancy
creamy
redistributing
fof
inflated
horace
indoctrinated
shuttleworth

```
In [240]: model_wikipedia200.most_similar('sparrow')

Out[240]: [('aim-7', 0.5671734809875488),
           ('saxaul', 0.5226368308067322),
           ('sparrows', 0.49139589071273804),
           ('grasshopper', 0.4559331238269806),
           ('warbler', 0.442558020335331726),
           ('starling', 0.43384942412376404),
           ('parrot', 0.4249793589115143),
           ('falcon', 0.4243379831314087),
           ('hummingbird', 0.4207715392112732),
           ('kestrel', 0.41665875911712646)]

In [241]: model_twitter200.most_similar('sparrow')

Out[241]: [('jack', 0.5756352543830872),
           ('kerouac', 0.489381343126297),
           ('pirate', 0.470211923122406),
           ('captain', 0.47001707553863525),
           ('depp', 0.46957796812057495),
           ('nicholson', 0.4578394591808319),
           ('nimble', 0.42101550102233887),
           ('johnny', 0.42048293352127075),
           ('pirates', 0.41835179924964905),
           ('picard', 0.4138476550579071)]
```

```
In [10]: model_wikipedia200.most_similar('girl')
```

```
Out[10]: [('boy', 0.8486549854278564),
          ('girls', 0.7696278691291809),
          ('woman', 0.7648226022720337),
          ('child', 0.7002282738685608),
          ('mother', 0.6969297528266907),
          ('teenage', 0.6899838447570801),
          ('boys', 0.6887997388839722),
          ('teen', 0.6872598528862),
          ('teenager', 0.6842571496963501),
          ('daughter', 0.6838234663009644)]
```

```
In [12]: model_twitter200.most_similar('girl')
```

```
Out[12]: [('boy', 0.8434211015701294),
          ('girls', 0.8288909792900085),
          ('she', 0.8030763864517212),
          ('guy', 0.7873061299324036),
          ('woman', 0.7817050218582153),
          ('chick', 0.7750226855278015),
          ('friend', 0.7702169418334961),
          ('bitch', 0.7611055374145508),
          ('that', 0.7493616342544556),
          ('pretty', 0.746584951877594)]
```

```
Twitter Clusters:

Cluster 1: ['politicking', 'trashing', 'butchering', 'archiving', 'fabric
ating', 'flaunting', 'cribbing']

Cluster 2: ['thirstiest', 'straightest', 'shortest', 'thickest', 'hippes
t', 'whitest', 'thinnest']

Cluster 3: ['induct', 'inducted', 'inducts', 'inducting', 'hof', 'inducti
ons', 'induction', 'inductees', 'inductee']

Cluster 4: ['bookshelf', 'shelving', 'bookcase', 'bookcases', 'headboar
d', 'shelves', 'bookshelves']

Cluster 5: ['conservatively', 'creatively', 'rationally', 'maturely', 're
phrase']

Cluster 6: ['one-off', 'half-hour', 'two-night', 'one-hour', 'two-hour']

Cluster 7: ['whole-heartedly', 'wholehearted', 'heartedly', 'heartily',
'respectfully', 'wholeheartedly', 'wholly']

Cluster 8: ['orthodontics', 'orthodontists', 'osteopath', 'orthodontist']
```

Wikipedia Clusters:

Cluster 1: ['twenty-six', 'seventy-five', 'thirty-five', 'forty-two', 'twenty-eight', 'twenty-nine', 'twenty-five', 'forty-five', 'four', 'thirteen', 'thirty-one', 'twenty-three', 'thirty-six', 'sixty-five', 'forty-six', 'sixty', 'three', 'forty', 'thirty', 'twenty-four', 'thirty-two', 'twenty-two', 'forty-eight', 'fourteen', 'sixty-four', 'thirty-three', 'twenty-one', 'thousand']

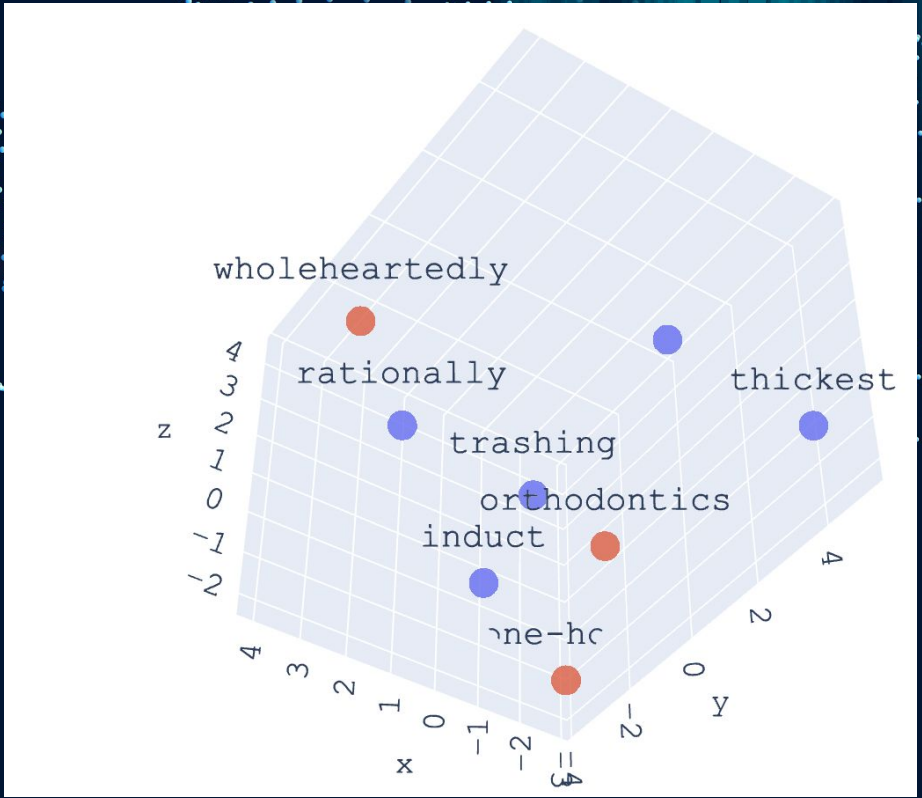
Cluster 2: ['exactly', 'really', 'surely', 'reason', 'importantly', 'though', 'actually', 'what', 'everything', 'fact', 'would', 'absolutely', 'thing', 'whoever', 'whatever', 'learned', 'although', 'somehow', 'think', 'this', 'things', 'ought', 'learn', 'undoubtedly', 'why', 'knew', 'everyone', 'might', 'know', 'thinks', 'nothing', 'certainly', 'hates', 'wherever', 'thought', 'whether', 'whenever', 'knowing', 'that', 'something', 'neither', 'thinking', 'clearly']

Cluster 3: ['warwickshire', 'northamptonshire', 'buckinghamshire', 'gloucestershire', 'monmouthshire', 'leicestershire', 'staffordshire', 'worcestershire', 'bedfordshire', 'northants', 'chorley', 'nottinghamshire', 'cheshire', 'oxfordshire', 'wiltshire', 'first-class', 'cambridgeshire', 'derbyshire', 'herefordshire', 'lancashire', 'lincolnshire', 'hertfordshire']

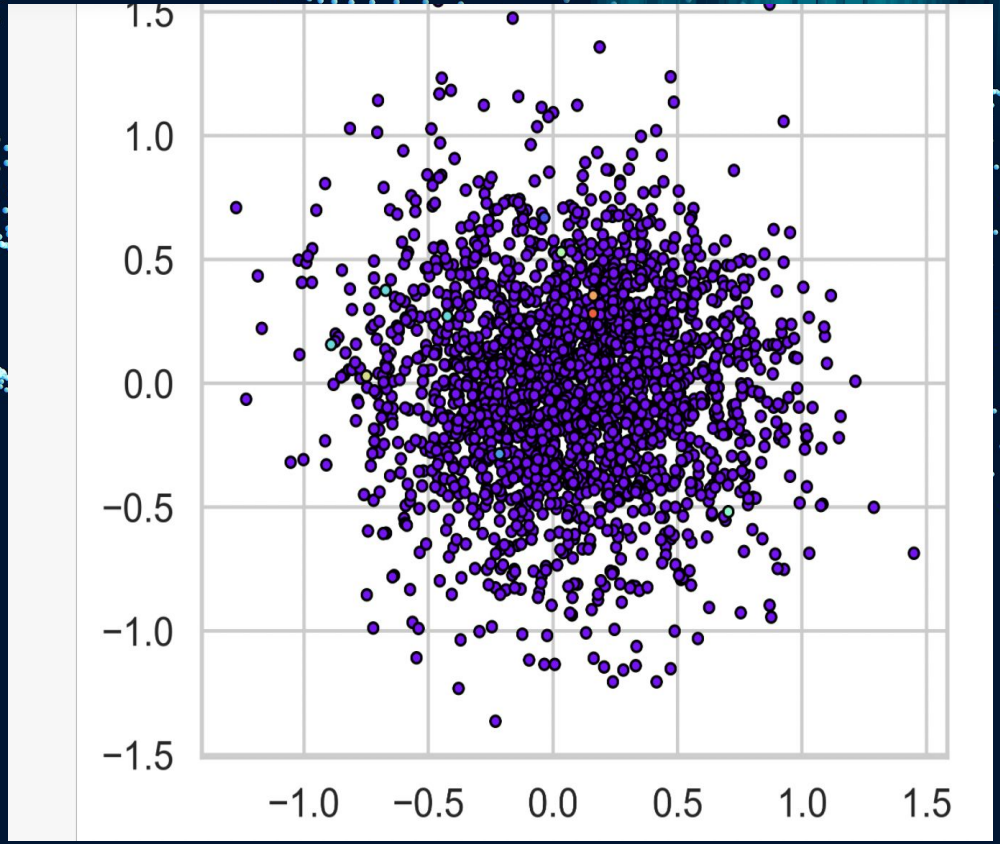
Cluster 4: ['fifth', 'twenty-first', 'finishing', 'thirteenth', 'straight', 'fourteenth', 'fifteenth', 'sixth', 'seventh', 'tenth', 'sixteenth', 'twelfth', 'fourth', 'first', 'ninth', 'third']

Cluster 5: ['father', 'brother', 'grandfather', 'nephew', 'his', 'stepfather', 'son-in-law', 'grandmother', 'brother-in-law', 'father-in-law', 'mother']

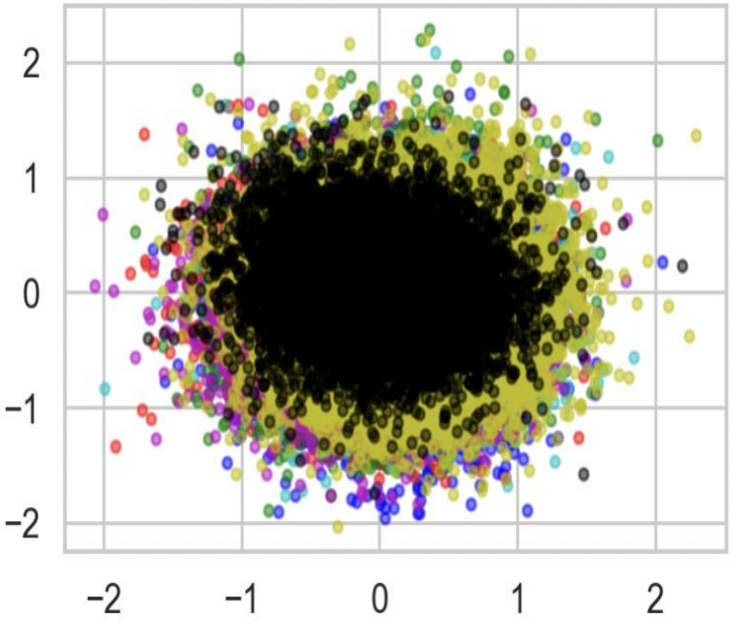
```
Cluster 6: ['indict', 'treason', 'pleading', 'plea', 'charge', 'implicate
d', 'counts', 'charging', 'charges', 'guilty', 'plead', 'arrested', 'plea
ded', 'theft', 'indicted', 'accused', 'convicts', 'convicted', 'suspect
s', 'pleads', 'convicting', 'pled', 'racketeering']

Cluster 7: ['northwestern', 'southwest', 'southern', 'east', 'southeaster
n', 'northeast', 'southwestern', 'northeastern', 'southeast', 'northern',
'northwest', 'eastern']

Cluster 8: ['forecasters', 'expecting', 'forecasted', 'predictions', 'exp
ected', 'prediction', 'expectations', 'predicted', 'foresaw', 'analysts',
'predicting', 'forecasting', 'forecasts', 'forecast', 'predicts']

Cluster 9: ['matches', 'champions', 'tournaments', 'match', 'tourney', 'c
hampionships', 'tournament', 'round', 'playoff', 'championship']

Cluster 10: ['catchers', 'homers', 'pitches', 'shortstop', 'pitching', 'c
atcher', 'hitters', 'pitchers', 'pitcher', 'hitter', 'pitched', 'diamondb
acks', 'fastball', 'lofton', 'leadoff']

Cluster 11: ['toasted', 'cooking', 'minced', 'roasting', 'broth', 'grate
d', 'eaten', 'chives', 'tofu', 'beans', 'onions', 'uncooked', 'cooked',
'spiced', 'chopped', 'finely', 'roasted', 'marinated', 'cook', 'onion',
'meat', 'celery']

Cluster 12: ['increased', 'higher', 'increase', 'reduce', 'increasing',
'reduction', 'reductions', 'increases', 'rates', 'reduced']

Cluster 13: ['shouts', 'shouted', 'crying', 'chanting', 'screaming', 'ins
ults', 'chanted', 'shout', 'screamed', 'shouting']
```

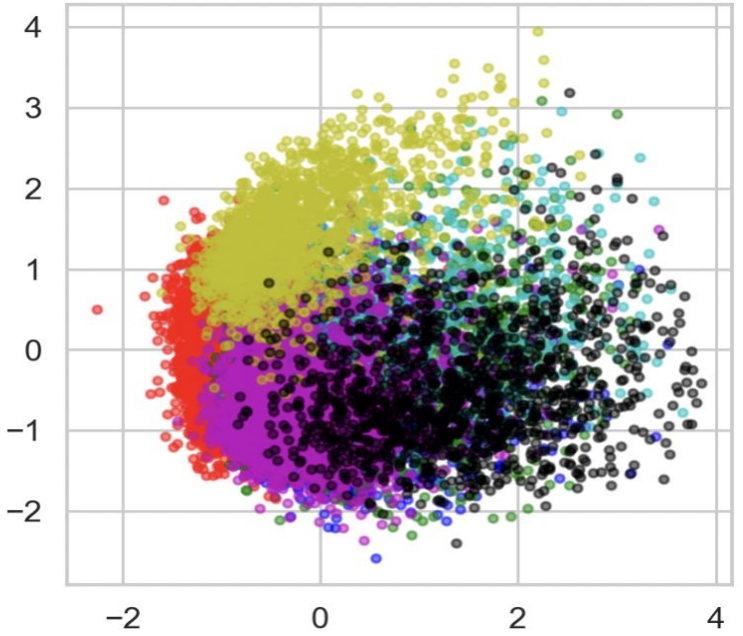


Twitter Clusters



Twitter DBSCAN - 125 Dimensions





K-Means Before and After Dimensionality Reduction