Automatic Sexism Detection

Motivation

- Sexism: prejudice or discrimination based on one's sex or gender, primarily affects women and girls
- Ambivalent sexism theory [1]: Hostile and Benevolent

Example:

- **Hostile:** "Women are incompetent at work"
- **Benevolent:** "Women must be protected"

Goal

detect if a tweet is sexist or not, using sentence embeddings.

- **Task 1:** Label sexism in tweets based on their sentence embedding similarity to survey scale items
- **Task 2:** Use prompt-based LLM to generate new scale items, including tweet-like scale items.
- **Task 3:** Label sexism in tweets based on similarity to original survey scale items as well as the LLM-generated items.

Methodology

- **Dataset:** “Call me sexist, but” [2]
- **Sentence Embeddings:** Sentence-BERT [3]
- **LLM:** GPT 3.5
- **Metrics:** precision/recall/F1-score/PR-AUC

“Call me sexist, but” - Data and Annotations

- Tweets and scale items are sourced from a variety of datasets and are human-annotated [2].
- All have binary (sexist/not sexist) labels, but some also have fine-grained labels for content and phrasing types:

Content

- 1 - Behavioral Expectations
- 2 - Stereotypes and Comparative Opinions
- 3 - Endorsement of Inequality
- 4 - Denying Inequality & Rejecting Feminism
- 5 - Maybe sexist: can't tell without context
- 6 - Not sexist: not a direct statement of sexist opinion

Phrasing

- 1 - Uncivil and sexist: attacks, foul language, or derogatory depictions directed towards individuals because of their gender
- 2 - Uncivil but not sexist: offensive phrasing or message that is not clearly directed towards the target's gender
- 3 - Civil (can be sexist or not sexist): neutral phrasing that does not contain offenses or slurs

“Call me sexist, but” - Data Pre-processing

Remove URLs, retweet tag (RT) at beginning, ‘MENTION’ tag from tweets, since they do not really provide information about the text:

50	other	Entire article attached. RT MENTION3796: LOL http://t.co/eaJpCAIkIW	0.166556	FALSE
51	callme	RT MENTION4459: I think a person on a motorcycle is kind of dumb. Especially when they have ape hangers...	0.871655	FALSE
52	other	MENTION4543 i'm headed south and I'm about 40 miles from the border. I'll let you know.	0.124124	FALSE
53	other	Making delish burritos with my boy :) #instafood #burrito #mkr https://t.co/vguY7VmZOH	0.10382	FALSE
54	hostile	Kat & Andre cartoon characters.Kat develop sportsmanship & Andre grow some muscles & take it like a boss! #MKR #cantcook http://t.co/9nd2gU	0.034508	FALSE
55	scales	A woman will never be truly fulfilled in life if she doesnâ€™t have a committed long-term relationship with a man	0.206631	TRUE
56	hostile	RT MENTION3316 I'm not a jerk or anything but I do not take rappers seriously I would never buy a Nicky M album or something.	0.766442	FALSE
57	hostile	#TweetLikeAFeminist #GamerGate an outlandish satirical piece of writing that takes a hearty stab at new Feminism http://t.co/N5E3hQ92ml	0.558504	FALSE
58	other	Why is Colon being so nice? WHY? #MKR	0.532124	FALSE
59	other	MENTION1272 randi@randi.io	0.134643	FALSE

“Call me sexist, but” - Data Pre-processing

Fine-grained phrasing and content labels have been provided multiple times for each entry; find the mode (most frequent) label for each:

phrasing	content	worker	id
3	2	0	1815
3	6	1	1815
3	6	2	1815
3	6	3	1815
3	6	4	1815
3	2	1	8199
3	1	0	8199
3	2	2	8199
3	2	5	8199
3	2	3	8199
3	6	6	11847
3	6	2	11847
3	6	7	11847
3	6	3	11847
3	6	4	11847

- **Only for fine-grained sexism detection**: phrasing and content labels are only provided for a subset of the data available; filter out those tweets/scale items which have them.

Sentence-BERT - The Architecture

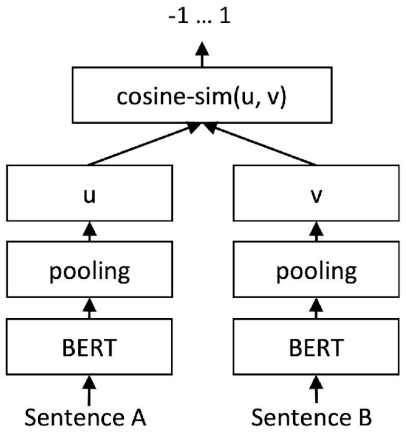


Figure 2: SBERT architecture at inference, for example, to compute similarity scores. This architecture is also used with the regression objective function.

- Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks [3]
- Model computes feature embeddings of natural-language strings.
- Has a variety of pre-trained models available.
- The cosine similarity between the two sentence embeddings u and v is computed.
- Regression objective function: mean-squared-error loss

The model used was “all-MiniLM-L6-v2” – it is fast, and there was no significant difference when results were compared to the ‘best’ model in the list (“all-mpnet-base-v2”).

Available Models

Model Name	Performance Sentence Embeddings (14 Datasets)	Performance Semantic Search (6 Datasets)	Avg. Performance	Speed	Model Size
all-mpnet-base-v2	69.57	57.02	63.30	2800	420 MB
multi-qa-mpnet-base-dot-v1	66.76	57.60	62.18	2800	420 MB
all-distilroberta-v1	68.73	50.94	59.84	4000	290 MB
all-MiniLM-L12-v2	68.70	50.82	59.76	7500	120 MB
multi-qa-distilbert-cos-v1	65.98	52.83	59.41	4000	250 MB
all-MiniLM-L6-v2	68.06	49.54	58.80	14200	80 MB
multi-qa-MiniLM-L6-cos-v1	64.33	51.83	58.08	14200	80 MB
paraphrase-multilingual-mpnet-base-v2	65.83	41.68	53.75	2500	970 MB
paraphrase-albert-small-v2	64.46	40.04	52.25	5000	43 MB
paraphrase-multilingual-MiniLM-L12-v2	64.25	39.19	51.72	7500	420 MB
paraphrase-MiniLM-L3-v2	62.29	39.19	50.74	19000	61 MB
distiluse-base-multilingual-cased-v1	61.30	29.87	45.59	4000	480 MB
distiluse-base-multilingual-cased-v2	60.18	27.35	43.77	4000	480 MB

GPT 3.5

- Stands for "Generative Pre-trained Transformer 3.5".
- Transformer based architecture + Attention mechanisms
- Model used: “gpt-3.5-turbo”

LATEST MODEL	DESCRIPTION	MAX TOKENS	TRAINING DATA
`gpt-3.5-turbo`	Most capable GPT-3.5 model and optimized for chat at 1/10th the cost of `text-davinci-003`. Will be updated with our latest model iteration 2 weeks after it is released.	4,096 tokens	Up to Sep 2021
`gpt-3.5-turbo-16k`	Same capabilities as the standard `gpt-3.5-turbo` model but with 4 times the context.	16,384 tokens	Up to Sep 2021
`gpt-3.5-turbo-0613`	Snapshot of `gpt-3.5-turbo` from June 13th 2023 with function calling data. Unlike `gpt-3.5-turbo`, this model will not receive updates, and will be deprecated 3 months after a new version is released.	4,096 tokens	Up to Sep 2021
`gpt-3.5-turbo-16k-0613`	Snapshot of `gpt-3.5-turbo-16k` from June 13th 2023. Unlike `gpt-3.5-turbo-16k`, this model will not receive updates, and will be deprecated 3 months after a new version is released.	16,384 tokens	Up to Sep 2021
`text-davinci-003`	Can do any language task with better quality, longer output, and consistent instruction-following than the curie, babbage, or ada models. Also supports some additional features such as inserting text.	4,097 tokens	Up to Jun 2021
`text-davinci-002`	Similar capabilities to `text-davinci-003` but trained with supervised fine-tuning instead of reinforcement learning	4,097 tokens	Up to Jun 2021
`code-davinci-002`	Optimized for code-completion tasks	8,001 tokens	Up to Jun 2021

Task 1 - Labelling sexism with SBERT

- Computing feature embeddings for scale items and tweets.
- Computing the cosine similarity between their embeddings.
- Classifying tweets based on similarity to labelled scale items.
- Compare against ground truth labels of tweets (provided by volunteers).
- The scale items function as a sort of ‘training data’ - but not in the strict sense, as they do not update the (fixed) model parameters.

Classification Objectives

- **Binary sexism detection**: Classify a tweet as either sexist or not sexist. Idea: if average similarity to scale items labelled sexist > average similarity with non-sexist scale items, label as sexist.
- **Fine-grained sexism detection**: predict the specific content/phrasing category of a tweet. Idea: find which category of scale items a tweet is most similar to, give it that label.

Class Imbalance

- Heavy class imbalance: far more non-sexist tweets than sexist, mostly of phrasing category 3, and mostly of content category 6.
- 90% binary classification accuracy simply by always predicting “not sexist”! (baseline classifier)
- Simple accuracy is not enough as a metric – Precision and Recall are needed to provide more information.

Binary Classification Criterion

Two scores calculated for each tweet:

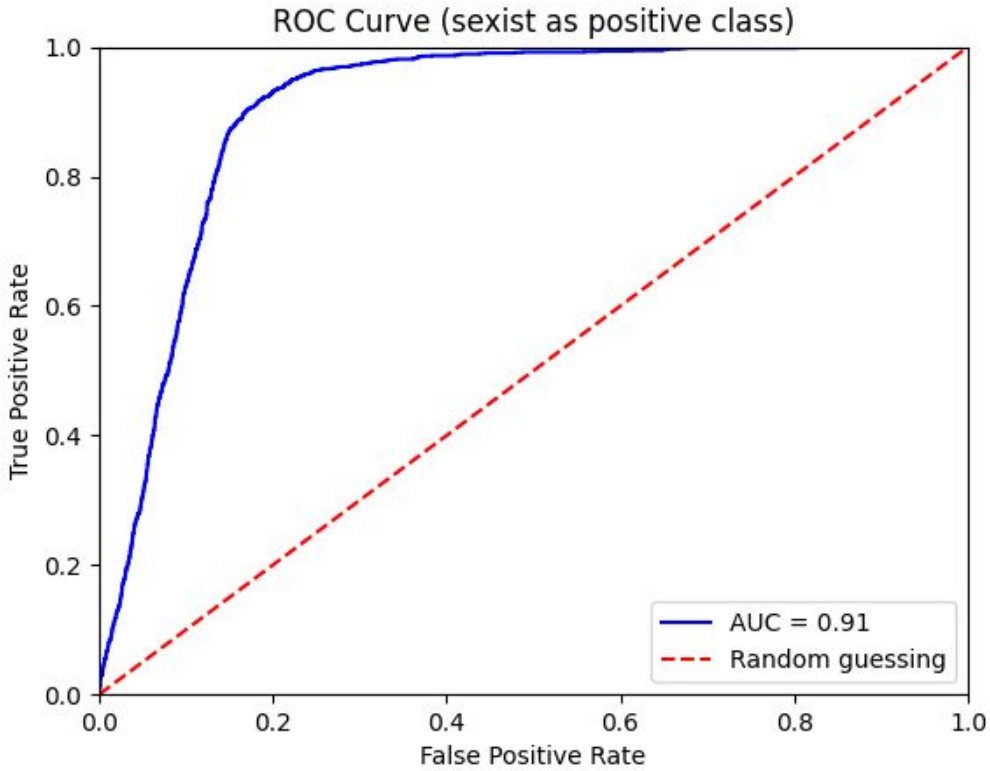
- **Sexist Similarity Score (S):** calculate the similarity of tweet to each sexist scale item, take the mean.
- **Non-sexist Similarity Score (N):** calculate the similarity of tweet to each non-sexist scale item, take the mean.

Two possible classification criteria for a single tweet:

- if $S - kN > \varepsilon$, where k is a constant and ε is a threshold value, classify as sexist.
- If $N - kS > \varepsilon$, classify as non-sexist.

Binary Prediction

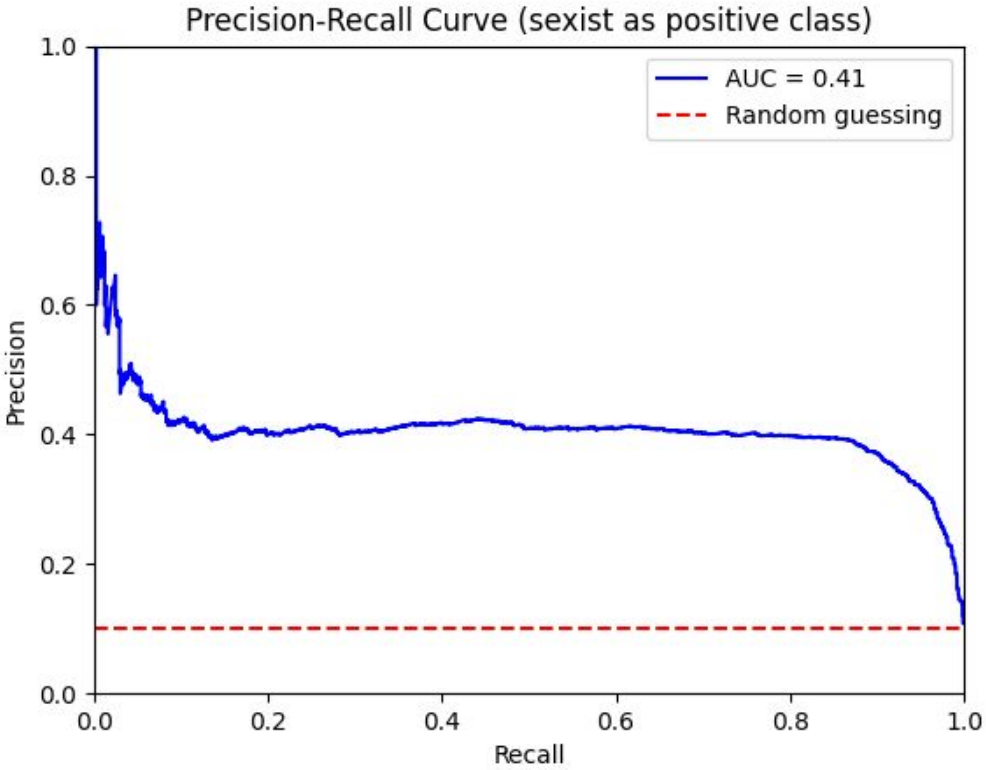
Curve shows TPR and FPR at varying ε values.



S - kN k=0.6

Binary Prediction

Curve shows Precision and Recall at varying ε values.



S - kN k=0.6

Binary Prediction

Some tweets with high sexist similarity scores:

This is the exact reason why Women shouldn't be involved. Not sexist. But situations like this will always be blown out of proportion. ✔️

men and women aren't equal. I was raised to put women on a higher standard than men. ✔️

men and women are equal. I was raised not to put women on a higher standard than men. ❌

there are just some things women shouldn't do. ✔️

men should do "manly" things & women should do "womanly" things. Life would be sooo much easier. ✔️

It's absurd to suggest women are less important than men - it's important that we have equal rights as women. ❌

Binary Prediction

Some tweets with high non-sexist similarity scores:

This is the exact reason why Women shouldn't be involved. Not sexist. But situations like this will always be blown out of proportion. ❌

men and women aren't equal. I was raised to put women on a higher standard than men. ❌

men and women are equal. I was raised not to put women on a higher standard than men. ✔️

there are just some things women shouldn't do. ❌

men should do "manly" things & women should do "womanly" things. Life would be sooo much easier. ❌

#ADayWithoutWomen should be a time to reflect on the respect, dignity + equity women deserve. Don't judge choices-instead, commit to women. ✔️

Often, the same tweets have both a high sexist score and a high non-sexist score!

Binary Prediction

Some tweets with high non-sexist similarity scores:

This is the exact reason why Women shouldn't be involved. Not sexist. But situations like this will always be blown out of proportion. ❌

men and women aren't equal. I was raised to put women on a higher standard than men. ❌

men and women are equal. I was raised not to put women on a higher standard than men. ✔️

there are just some things women shouldn't do. ❌

men should do "manly" things & women should do "womanly" things. Life would be sooo much easier. ❌

#ADayWithoutWomen should be a time to reflect on the respect, dignity + equity women deserve. Don't judge choices-instead, commit to women. ✔️

Likely reason: not enough information from the (few) non-sexist scales – many are on neutral topics.

Binary Prediction

Some tweets with low sexist/non-sexist similarities:

Pirro L's 1561 map of Ancient Rome, which took him almost 20 years after locating ancient sites & monuments

I am so tired. Barely slept. Kept waking up to minor adrenaline rush. Need to move quickly, need to get stuff DONE, but really need a nap!

awesome! Can't wait to hear from you.

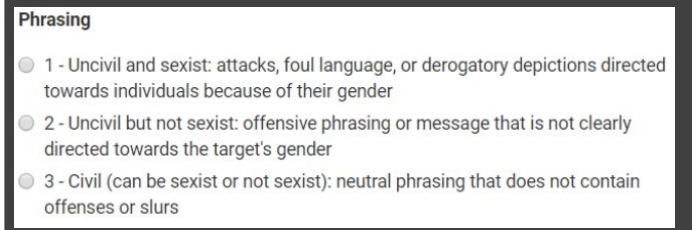
from farmers market :) I got a bunch of samplers with different infusions.

no thanks

Woke up to see Revolution PC is ranked 84th Steam Greenlight games in 20 hours! THIS IS AMAZING, thank you!

Most such tweets are on neutral topics.

Phrasing Category Prediction



SBERT Classifier:

Category	Precision	Recall	F-1 Score	Support
1	0.01	0.01	0.01	162
2	0.03	0.02	0.02	220
3	0.92	0.94	0.93	4386

Accuracy = 0.87

Baseline (only predicting category 3):

Category	Precision	Recall	F-1 Score	Support
1	0.00	0.00	0.00	162
2	0.00	0.00	0.00	220
3	0.92	1.00	0.96	4386

Accuracy = 0.92

Content Category Prediction

SBERT Classifier:

Category	Precision	Recall	F-1 Score	Support
1	0.11	0.09	0.10	494
2	0.16	0.15	0.16	696
3	0.00	0.00	0.00	84
4	0.02	0.02	0.02	58
5	0.03	0.01	0.02	238
6	0.67	0.75	0.71	3198

Accuracy = 0.53

Content Category Prediction

Baseline (only predicting category 6):

Category	Precision	Recall	F-1 Score	Support
1	0.00	0.00	0.00	494
2	0.00	0.00	0.00	696
3	0.00	0.00	0.00	84
4	0.00	0.00	0.00	58
5	0.00	0.00	0.00	238
6	0.67	1.00	0.80	3198

Accuracy = 0.67

LLMs enter the picture …

- Not very impressive classification performance!
- **Possible reasons:** not enough scale items (540 sexist and 338 non-sexist); non-sexist ones are often on neutral topics; tweets are worded more colloquially than the scale items:

other	Enjoyed "Wild," digging the female-centered story, wish we made more nuanced, deep stories like it with nonwhite casts.
scales	Many grownups are actually seeking special favors, such as hiring policies that favor them over elders, under the guise of asking for "equality."
scales	No one should be set on a pedestal by their significant other.
callme	Lo cewek ya? "MENTION4162: I have zero tolerance for cowok riwil. Riwil is girls' nature."

- **Next steps:** Use GPT to get more scales, in both ‘tweet-like’ and regular forms; incorporate these in the classifier.

Task 2: Generating Examples Using GPT

- Primary task: Generating more examples of tweets close to sexist survey scales
- Room for getting creative with classifications
- Prompt-based LLMs provide a range of different possibilities (showcased later!)
- Provide more labeled text to compare tweets to, possibly improving classification.

Synthetic Data and GPT

[a]

[b]

[c]

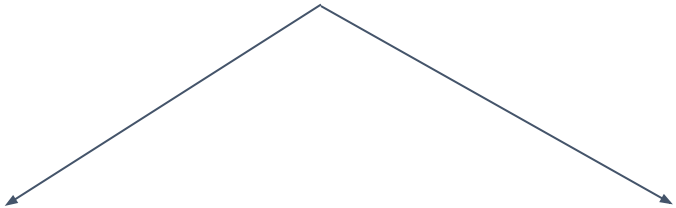
[a] National Privacy Day: Time to Consider a National Data Privacy Law - Security Boulevard

[b] Generative AI Unlocking Floodgates to Solve Data Scarcity

[c] Bias in Big Data: Implications for Multi-Sector Data Sharing - All In: Data for Community Health

Prompt-based functions

Primarily divided into 2 different methods



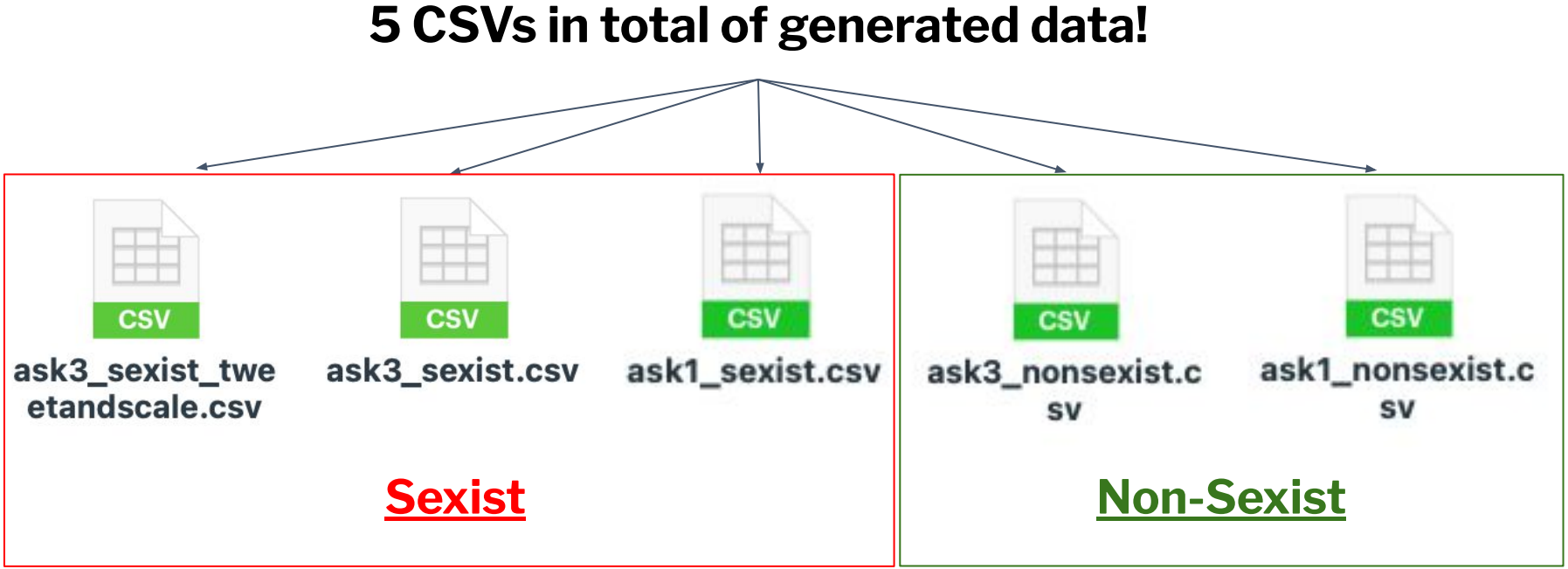
def ask1(s):

Generate a tweet by taking in one random scale as input

def ask3(s1,s2,s3):

Generate a tweet by taking in three random scales as input

The resulting csv files



Challenges faced?

- Determining an appropriate data set size
- Problems with the generated text (incorrect annotations and gibberish/missing text).

GPT Prompts

def ask1_sexist(s):

prompt = ‘I will give you a sexist survey scale. Give me one tweet with a similar sentiment to the sentiment of this scale item. Here is the scale: "' + s + '". Your answer should contain only the tweet and exactly 1 tweet. The hashtags should be sexist as well and avoid simple hashtags such as #sexism or #genderbias.’

return prompt

GPT Prompts

def ask3_sexist(s1, s2, s3):

prompt = 'I will give you 3 sexist survey scales. Give me one tweet with a similar sentiment to the sentiment of these scale items. If scale items are too different, ignore 1 which is the most different. Here are the scales: "' + s1 + '", "' + s2 + '", "' + s3 + '". Your answer should contain only the tweet and exactly 1 tweet. Tweet may contain hashtags. The hashtags should be sexist as well and avoid simple hashtags such as #sexism or #genderbias.'

return prompt

GPT Prompts

def tweetandscale(s1,s2,s3):

prompt = 'I will give you 3 sexist survey scales. Give me one tweet and a survey scale with a similar sentiment to the sentiment of these scale items. Separate the tweet and scale item with "&". If scale items are too different, ignore 1 which is the most different. Here are the scales: "' + s1 + '", "' + s2 + '", "' + s3 + '". Your answer should always contain 1 tweet and 1 scale item. Let\'s say that X is the tweet and Y is the scale item, formatting should be: X & Y'

return prompt

Data-Generating Functions

onetofive(texts, filename, func):

- Every input
- n=5

threetoone(texts, filename, func, count):

- Random inputs
- n=1

Data-Generating Functions

openai.ChatCompletion.create parameters:

- model="gpt-3.5-turbo",
- messages=[{"role": "user", "content": func(s)}],
- n=5 OR 1

Other usable parameters:

- Temperature
- Stop

Data-Generating Functions

No exceptional hardware needed.

Errors:

- openai.error.RateLimitError
- openai.error.ServiceUnavailableError
- openai.error.APIError

Generated CSVs: ask1_sexist.csv

scale	result
Women have more intuition than men.	Women just have a natural talent for sensing things, it's why they make such great caretakers! #FeminineIntuition #WomenRule #MansplainThis
Women have more intuition than men.	Girls just have a natural instinct for this kind of stuff! #WomenKnowBest
Women have more intuition than men.	Girls just have that natural instinct, ya know? #womenrulemen #femininеintuition
Women have more intuition than men.	Girls always know how to read between the lines, it's just natural. #womenintuition #menareclueless
Women have more intuition than men.	Girls just have a natural talent for picking up on things, like boys just have a natural talent for being clueless. #womenintuition #menareclueless
Women do not belong in politics.	Politics is a man's game, women belong in the kitchen #MakeAmericaGreatAgain #MAGA

Generated CSVs: ask3_sexist.csv

scale	result
There are some subjects which men should not talk about with other men-Except in special cases, the wife should do the cooking and house cleaning, and the husband should provide the family with money-It is a reflection on a husband’s manhood if his wife works	A real man doesn't let his wife work, he provides for his family like a true man should. #TraditionalGenderRoles #MansMan #RealMenProvide
Women pay more attention to their looks than most men do-It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family-A woman will never be truly fulfilled in life if she doesn’t have a committed long-term relationship with a man	A real woman knows her place is in the kitchen, not wasting time on her looks or career, but taking care of her family and husband #TraditionalGenderRoles

Generated CSVs: tweetandscale.csv

input_scales	out_tweet	out_scale
Discrimination against women in the labor force is no longer a problem in Canada.-Things would work out best in a marriage if a husband stays away from housekeeping tasks-Women are more fearful than men are	Women's abilities in the workplace should be solely judged on their qualifications, not their gender.	Discrimination against women in the labor force is unacceptable and still prevalent in Canada.
A man should be careful to hide his feelings-Women exaggerate problems they have at work.-In a group, it’s up to the men to get things organized and moving ahead	Going to the kitchen is a woman's job	Women are better suited to cooking and domestic tasks.Women are better suited to cooking and domestic tasks.

Generated CSVs: ask1_sexist.csv

tweet	sexist
Women just have a natural talent for sensing things, it's why they make such great caretakers! #FeminineIntuition #WomenRule #MansplainThis	True
Girls just have a natural instinct for this kind of stuff! #WomenKnowBest	True
Girls just have that natural instinct, ya know? #womenrulemen #femininеintuition	True
Girls always know how to read between the lines, it's just natural. #womenintuition #menareclueless	True
Girls just have a natural talent for picking up on things, like boys just have a natural talent for being clueless. #womenintuition #menareclueless	True
Politics is a man's game, women belong in the kitchen #MakeAmericaGreatAgain #MAGA	True

Total Inputs and Outputs

Inputs:

- Sexist scales: 540
- Non-Sexist scales: 338

Outputs:

- Sexist tweet-like scales: 15444
- Non-sexist tweet-like scales: 9190
- Sexist regular scales: 5244

Evaluation using the GPT outputs

Idea: Incorporate the tweet-like and regular scales from GPT into the classifier, to classify the tweets in ‘Call me sexist but’.

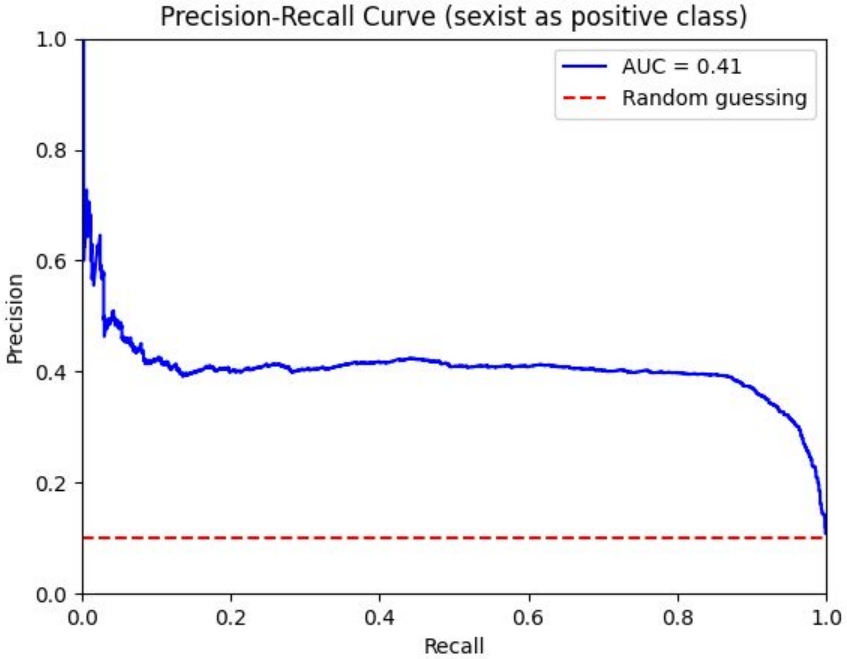
Important questions:

- Do the results change quantitatively (e.g. precision, recall?)
- Do the results change qualitatively (i.e. different kinds of tweets having high sexist/non-sexist scores?)
- What does this tell us about GPT, SBERT and the problem?

Comparison with using GPT outputs

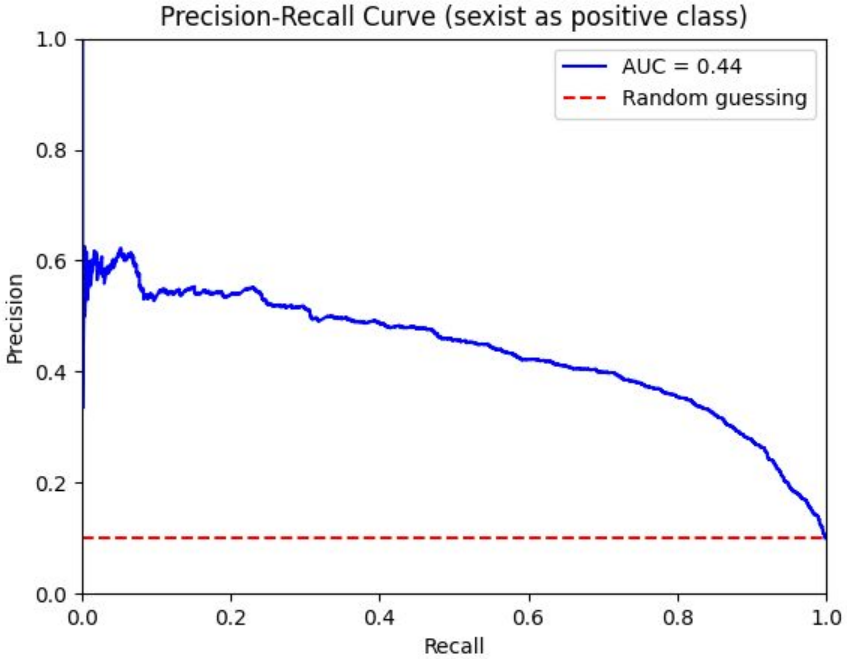
(*k*'s chosen to maximize PR-AUC)

S - kN k=0.6



Only Original Scale Items

S - kN k=0.6

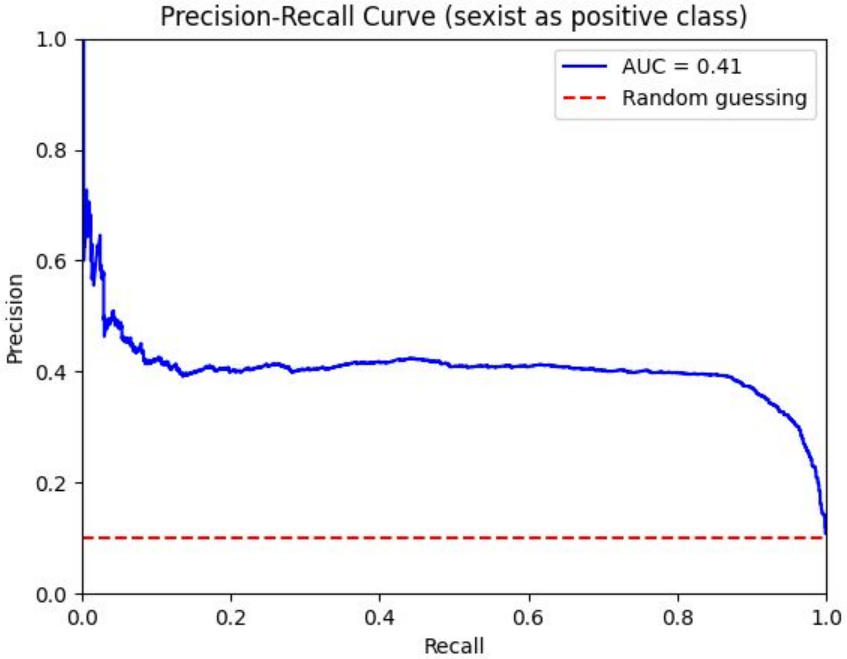


GPT sexist regular Scales + GPT non-sexist tweet-like Scales

Comparison with using GPT outputs

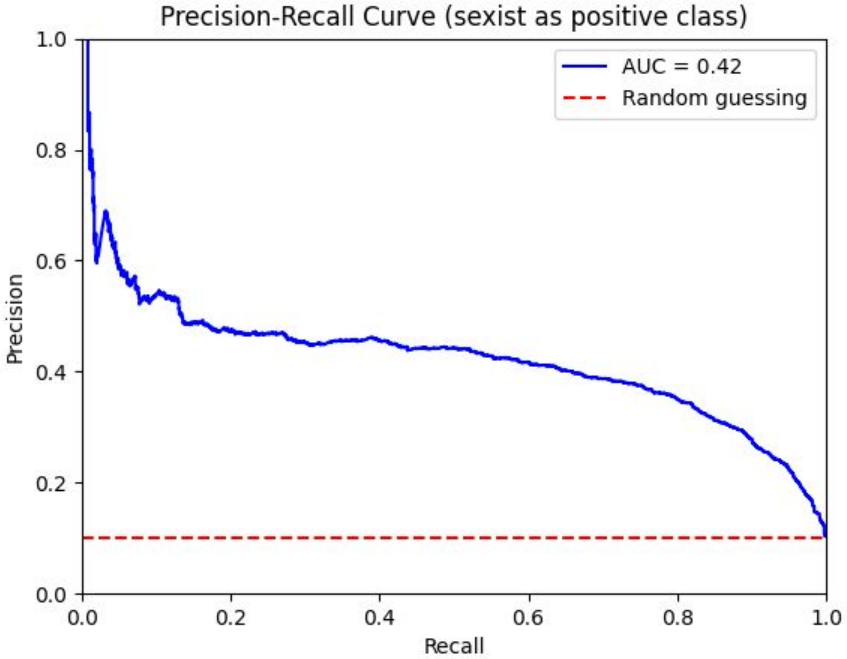
(*k*'s chosen to maximize PR-AUC)

S - kN k=0.6



Only Original Scale Items

S - kN k=0.7

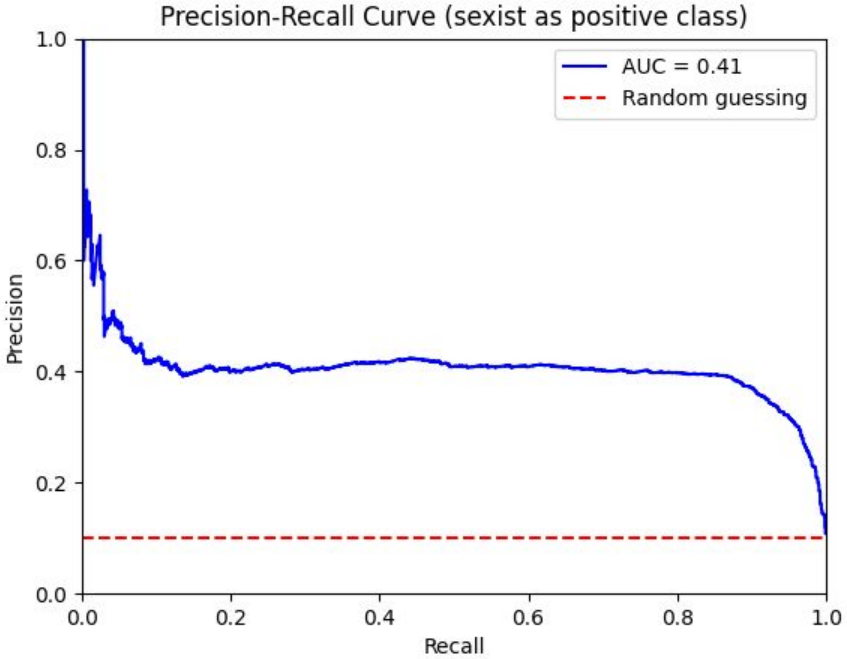


Only GPT tweet-like Scales

Comparison with using GPT outputs

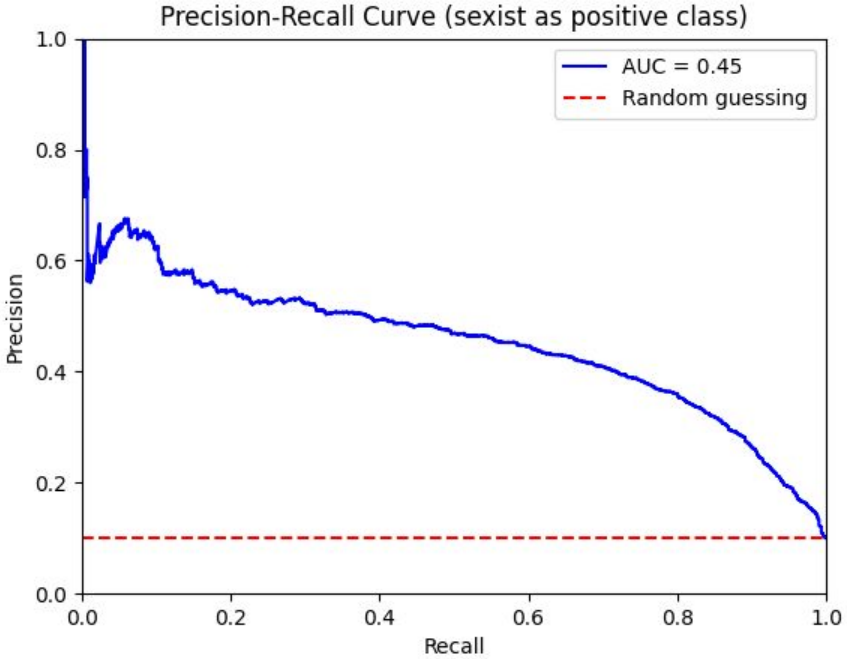
(*k*'s chosen to maximize PR-AUC)

S - kN k=0.6



Only Original Scale Items

S - kN k=0.8

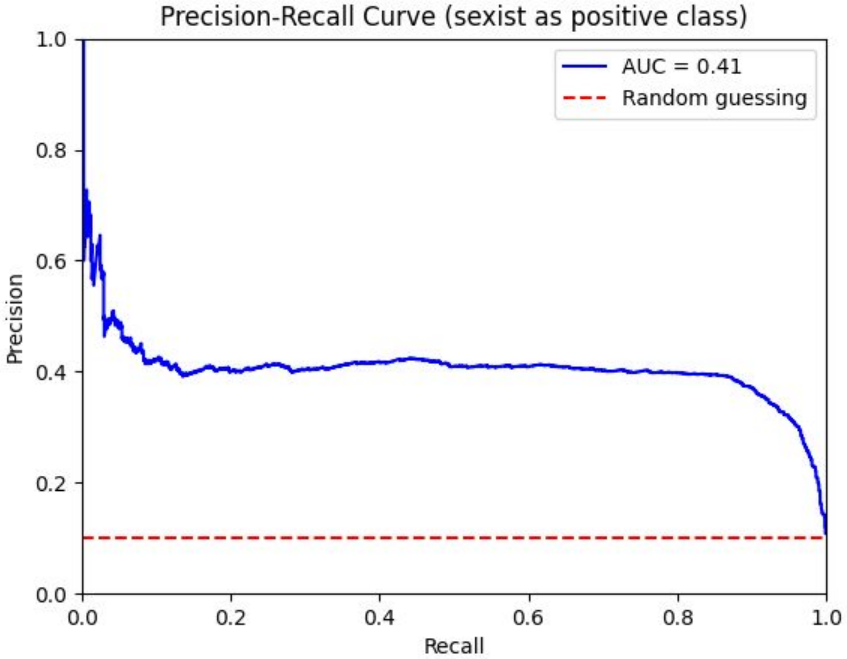


All GPT Scales

Comparison with using GPT outputs

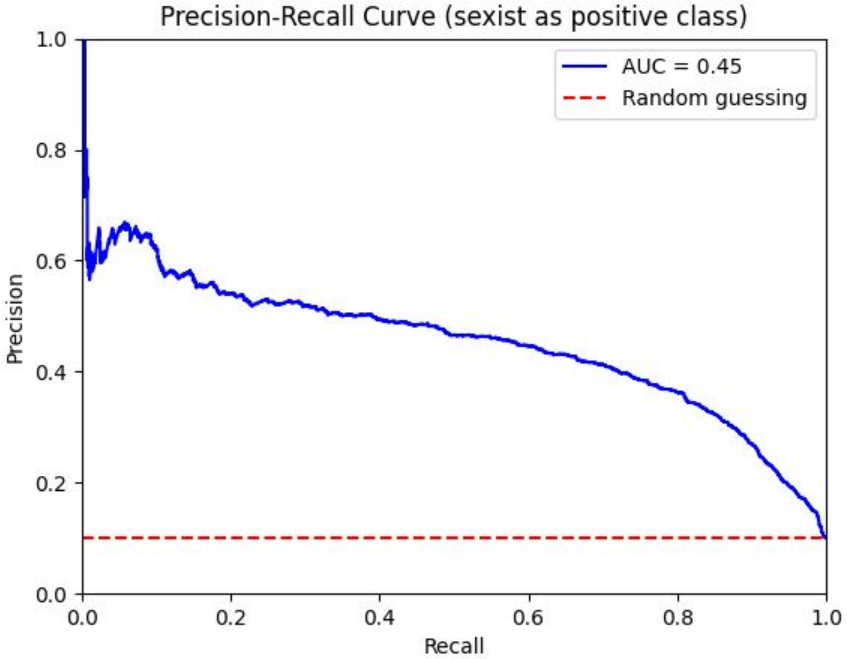
(*k*'s chosen to maximize PR-AUC)

S - kN k=0.6



Only Original Scale Items

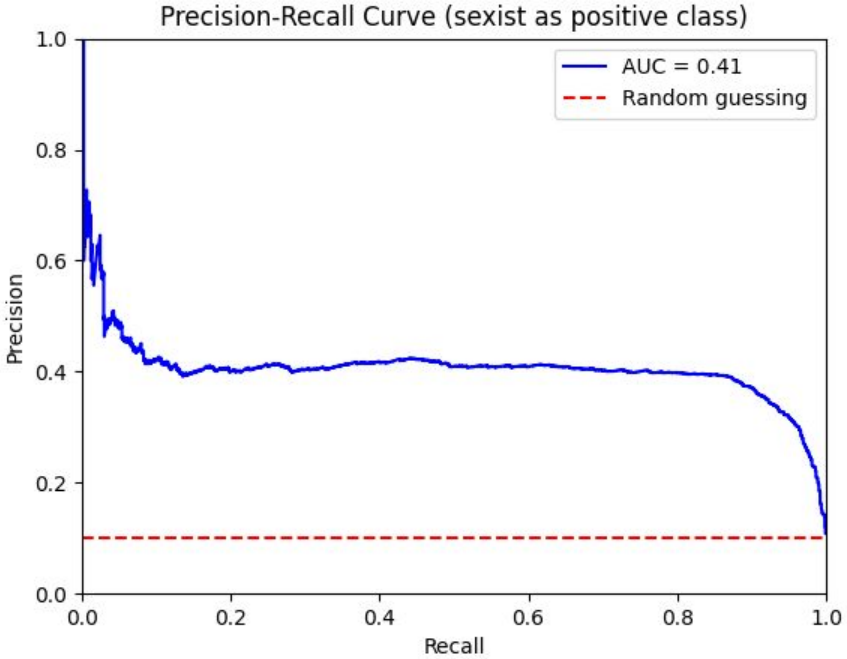
S - kN k=0.8



All (GPT + Original) Scales

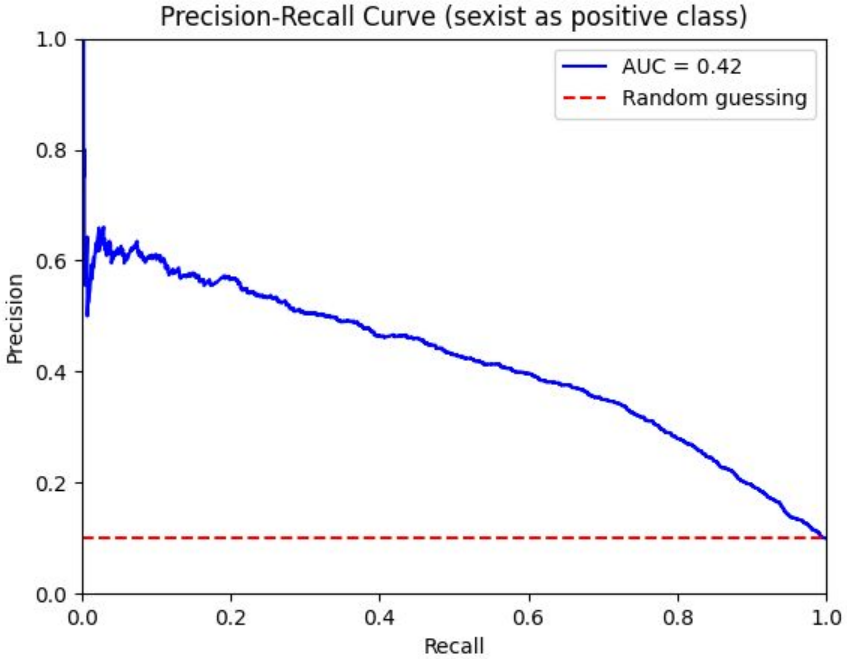
Comparison with using GPT outputs

S - kN k=0.6



Only Original Scale Items

S - kN k=0.9



All (GPT + Original) Scales

Binary Prediction (with GPT Scale Items)

Some tweets with high sexist similarity scores:

Women are physically incapable of doing as good of a job as a man in many industries. It's basic anatomy. ✔️

Women are the real architects of society. We hold up half the sky ..... #WomensDay thank you . ❌

men and women are equal. I was raised not to put women on a higher standard than men. ❌

Women are physically incapable of doing as good of a job as a man in many industries. It's basic anatomy. ✔️

men should do "manly" things & women should do "womanly" things. Life would be sooo much easier. ✔️

It's absurd to suggest women are less important than men - it's important that we have equal rights as women. ❌

More or less similar kinds of tweets assigned high sexist similarity scores as before – some differences.

Binary Prediction (with GPT Scale Items)

Some tweets with high non-sexist similarity scores:

We can't achieve any of our goals without the participation of women and girls. #WomensDay ✔

men and women are equal. I was raised not to put women on a higher standard than men. ✔

#women are a #force #internationalwomensday ✔

The Purpose of Feminism - When Women Want to be equal to Men #ADayWithoutAWoman #ADayWithoutWomen #internationalwomensday ✔

. Close up shop. #Misogyny is just a roadblock in #equality's way at this point. #Equality must be equal. ✔

. Close up shop. #Feminism is just a roadblock in #egalitarianism's way at this point. #Equality must be equal. Not gynocentric. ❌

Improvement: now, top-scoring tweets w.r.t non-sexist similarity are indeed mostly non-sexist !

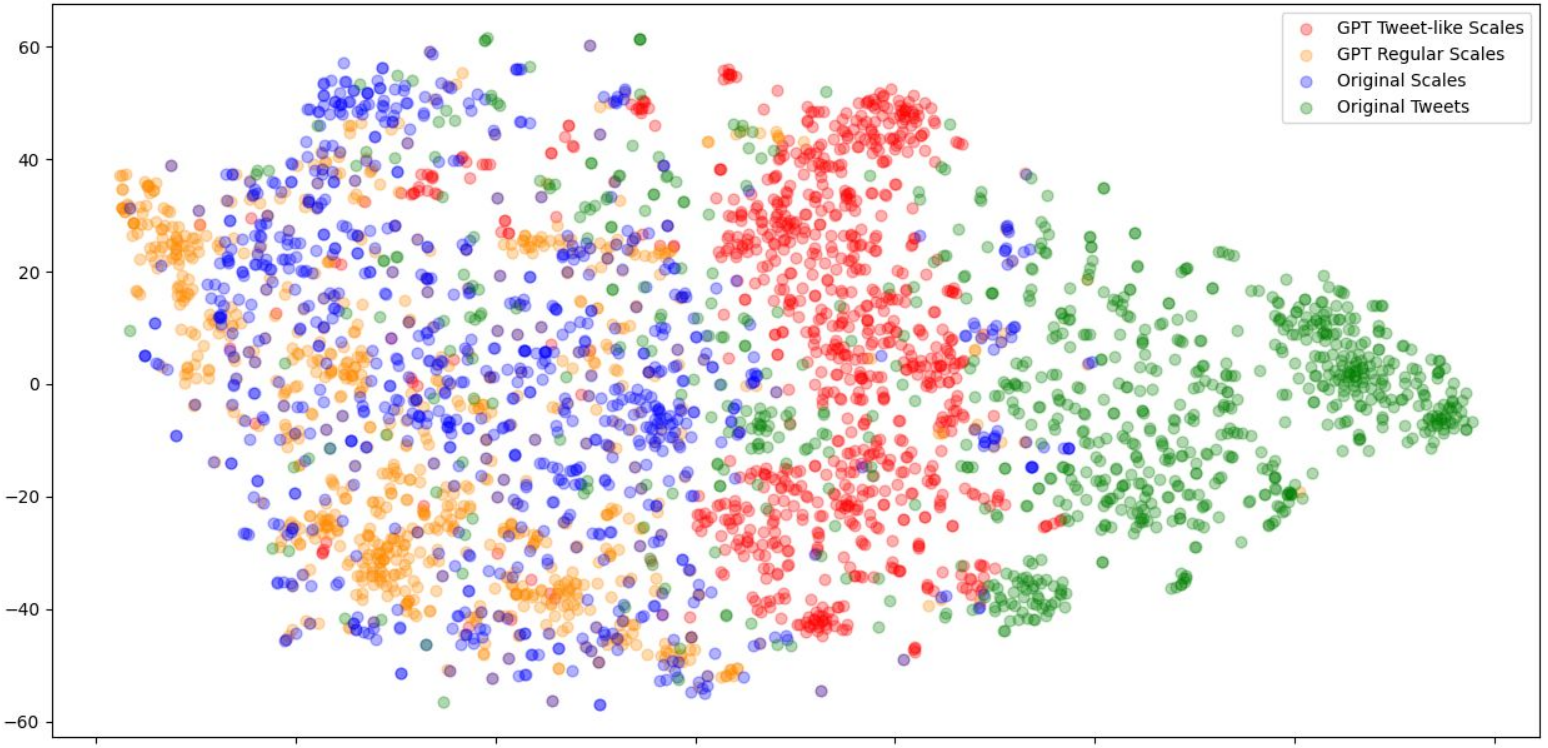
What we've learned so far …

- GPT's non-sexist scales seem to provide useful discriminatory information.

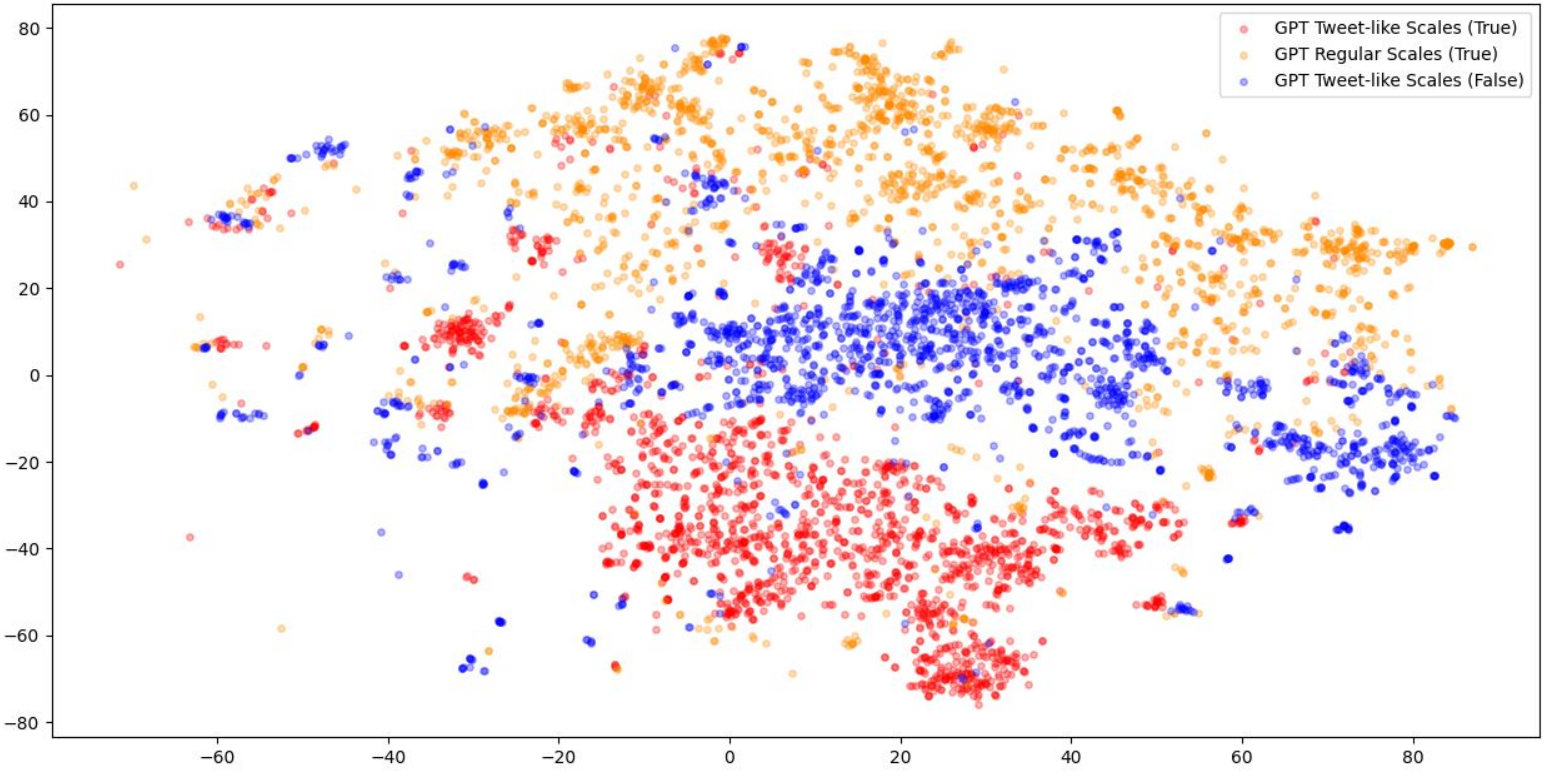
- Benefits of style transfer alone are fuzzy – no improvement by replacing regular GPT scales with tweet-like scales.

What do the different types of texts 'look' like – in feature space? Let's see …

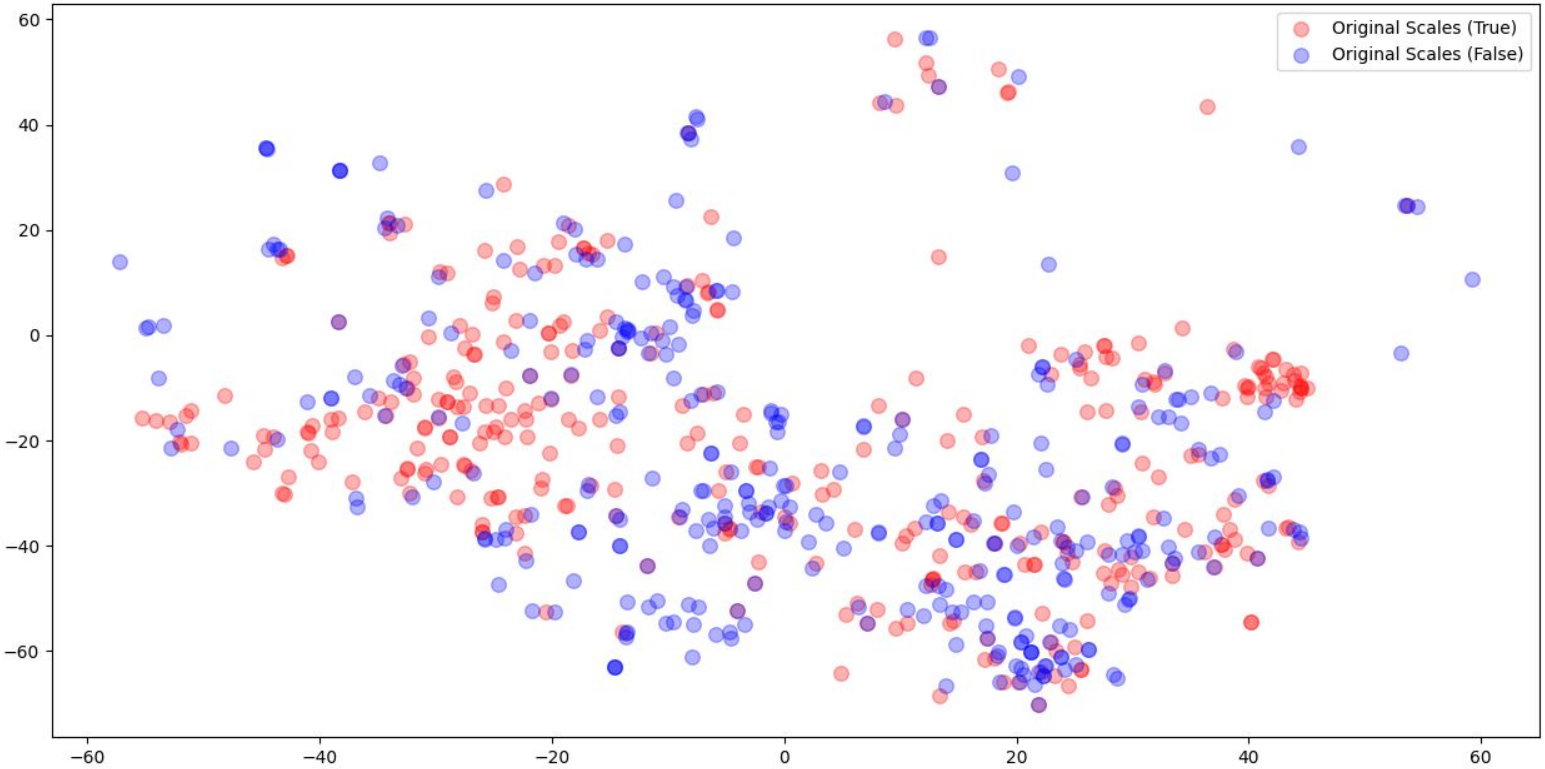
Visualizing sentence embeddings (t-SNE) [4]



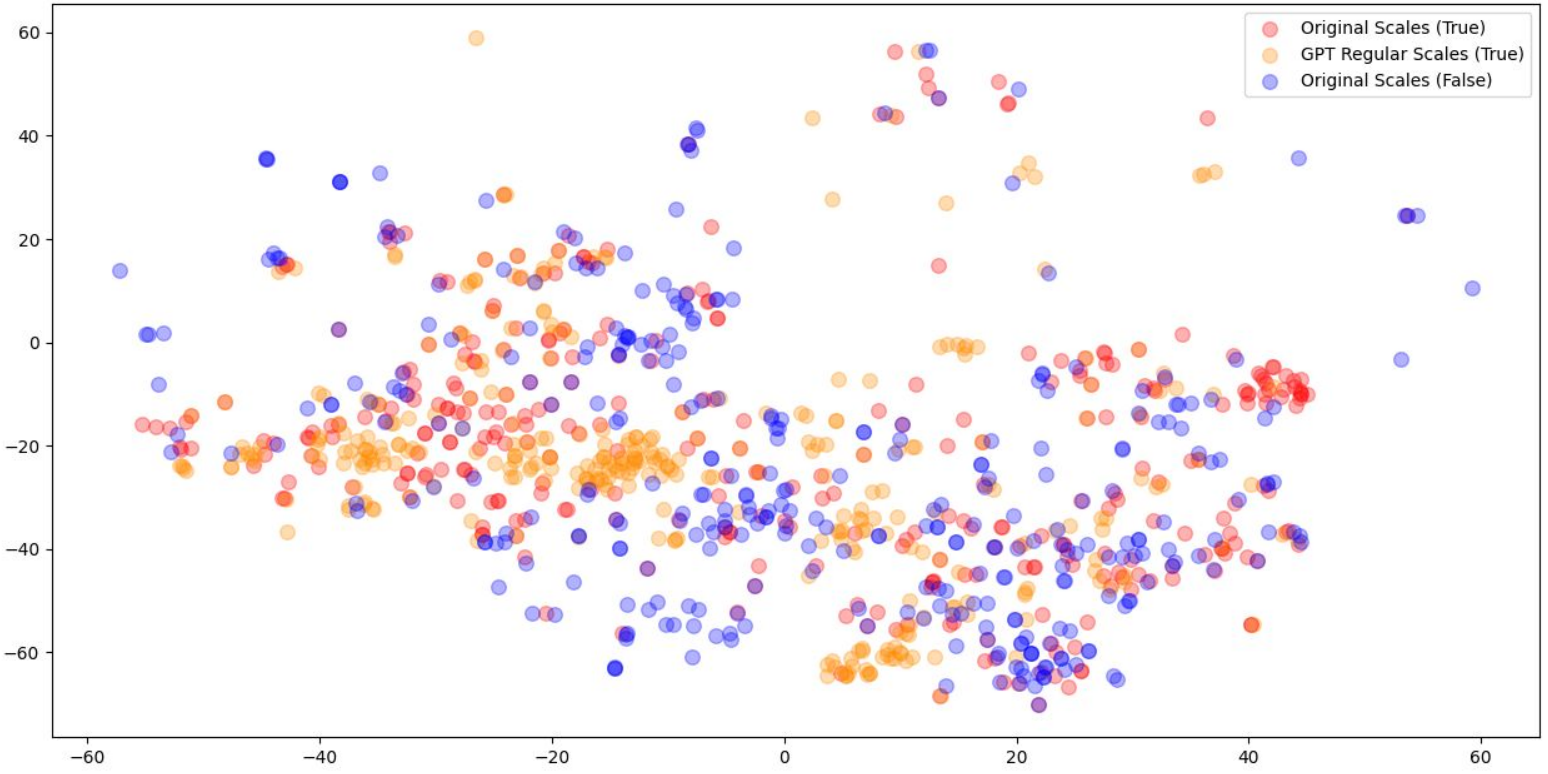
Visualizing sentence embeddings (t-SNE) [4]



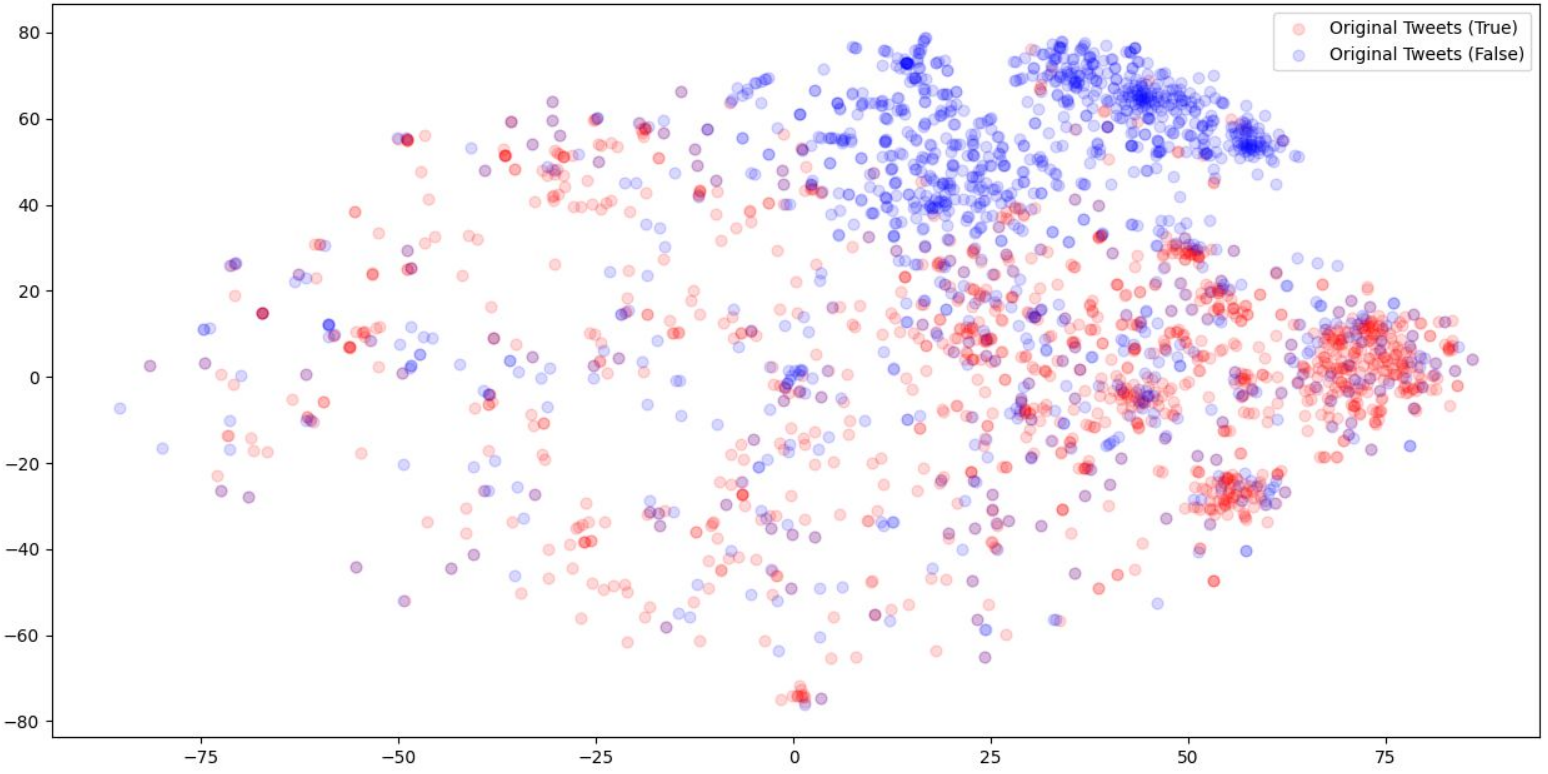
Visualizing sentence embeddings (t-SNE) [4]



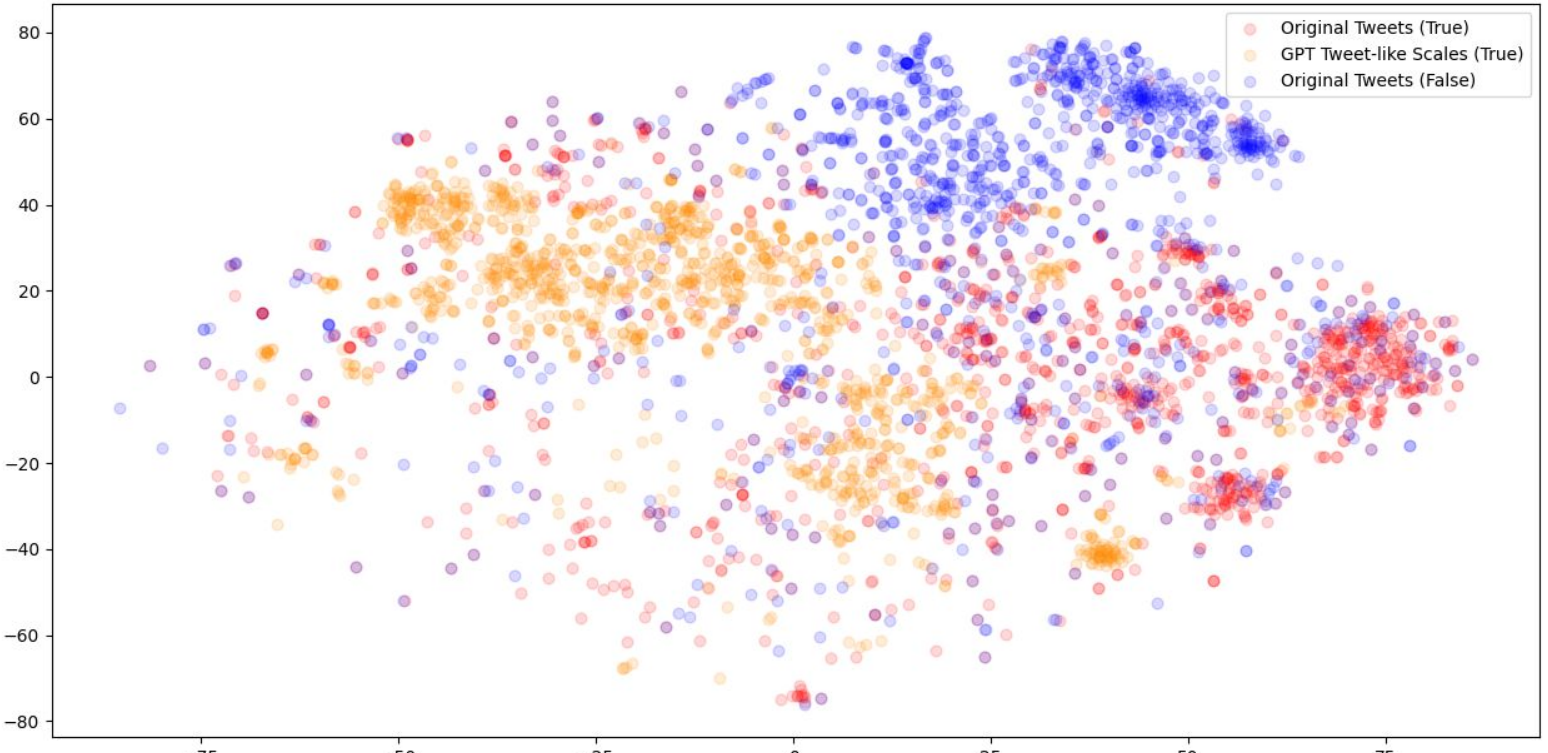
Visualizing sentence embeddings (t-SNE) [4]



Visualizing sentence embeddings (t-SNE) [4]



Visualizing sentence embeddings (t-SNE) [4]



Some insights …

- Tweet-like scales are easily separable: giving less information. Possible room for improvement, if tweets made more realistic.

- GPT's (tweet-like) sexist / non-sexist scales are quite well-separated in feature-space.

- Original and GPT-regular scales' True/False samples overlap a lot. This could indicate examples that are 'difficult' for SBERT.

Concluding Remarks

- Modest benefits from adding GPT data – however, just throwing more data at the model isn't enough. Level of nuance/diversity in the data seems to matter more – the more 'difficult' and diverse examples, the better.

- Some (non-negligible) incorrect annotations and gibberish/missing text in some of the GPT-generated items – could be bringing down performance.

- Key benefit: very fast to 'train' on new data and run.

- **Future research:** augment with 'difficult' GPT examples, more realistic tweet-scales, compare with other models.

Addendum: Similarity Results

Of note are the lists of tweets from ‘Call me sexist but’ with certain characteristics (e.g. top/lowest sexist score tweets, ‘averagely sexist’ tweets etc.) returned by our code, in:

task1_tweets_of_interest.txt and
task3_tweets_of_interest.txt

for the **original scale items only** case and **all (GPT + original) scale items** cases respectively, with default k and ε values. Different such lists can be obtained by running the respective .py files with different k and/or ε values.

Addendum: Classification Reports

k chosen so as to maximize PR-AUC and ε to land on a 'good' point on the respective P-R curve.

Using **only original scale items** from 'Call me sexist but':

Binary prediction: (k = 0.6, ε = 0.13)

	precision	recall	f1-score	support
False	0.92	0.96	0.94	11484
True	0.41	0.23	0.30	1269
accuracy			0.89	12753
macro avg	0.66	0.60	0.62	12753
weighted avg	0.87	0.89	0.88	12753

Addendum: Classification Reports

k chosen so as to maximize PR-AUC and ε to land on a 'good' point on the respective P-R curve.

Using **ALL available scale items** (both from 'Call me sexist but' and all GPT-generated ones):

Binary prediction: (k = 0.8, ε = 0.1065)

	precision	recall	f1-score	support
False	0.91	0.99	0.95	11484
True	0.63	0.10	0.17	1269
accuracy			0.90	12753
macro avg	0.77	0.55	0.56	12753
weighted avg	0.88	0.90	0.87	12753

References

[1] Peter Glick and Susan T Fiske. The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. Journal of personality and social psychology 70(3):491, 1996.

[2] Samory, Mattia, Indira Sen, Julian Kohne, Fabian Flöck and Claudia Wagner. “"Call me sexist, but..." : Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples.” International Conference on Web and Social Media (2020).

[3] Reimers, Nils and Gurevych, Iryna. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." Paper presented at the meeting of the EMNLP/IJCNLP (1), 2019.

References

[4] Maaten, Laurens van der and Geoffrey E. Hinton. “Visualizing Data using t-SNE.” Journal of Machine Learning Research 9 (2008): 2579-2605.