Developing a Method for the Disentanglement of Style from Content in Textual Data

Presented by: Michelle Fribance, M.Sc.

Principal supervisor: Prof. Dr. Mathias Kraus

Associate supervisor: Nico Hambauer, M.Sc.

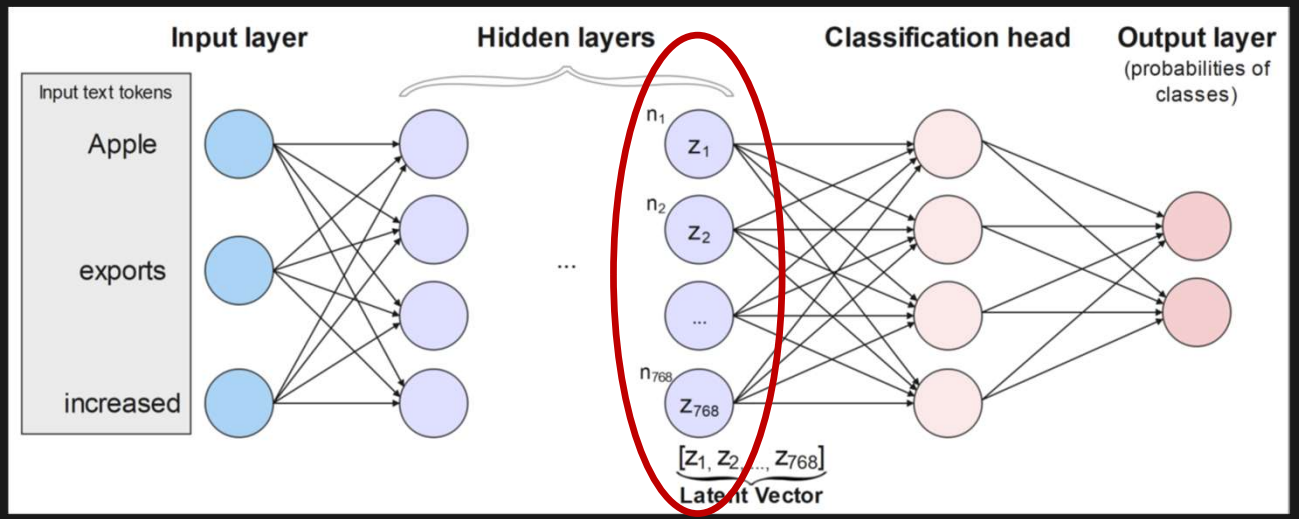
Topic Intro

Goal: Enhancing the explainability of ‘Black Box’ neural networks & improving model performance.

Method: Disentanglement of the latent vector space in the embeddings of the network’s hidden layers.

Focus: Post-processing of a pretrained NLP model by isolating style from content factors.

Latent Vectors

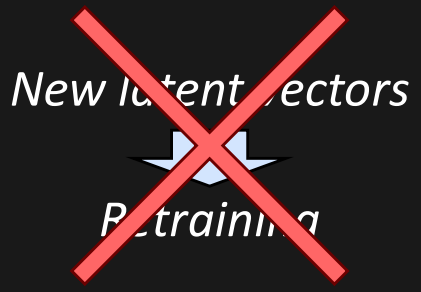


Disentanglement

Separating out the main factors of variation present in a data distribution.

Existing methods:

- Pre-processing methods: High computational & time costs → limited scalability
- Existing post-processing methods: Primarily computer vision-based
- Existing NLP style vs. content methods: Single stylization



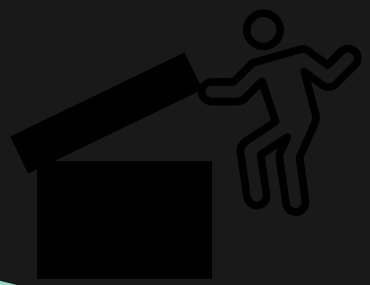
Explainable AI

Simplifying the connection between the outcome of a model and the input.

input ? *output*

SHAP
LIME
Surrogate models

Improving the understandability of the inner workings of a model.



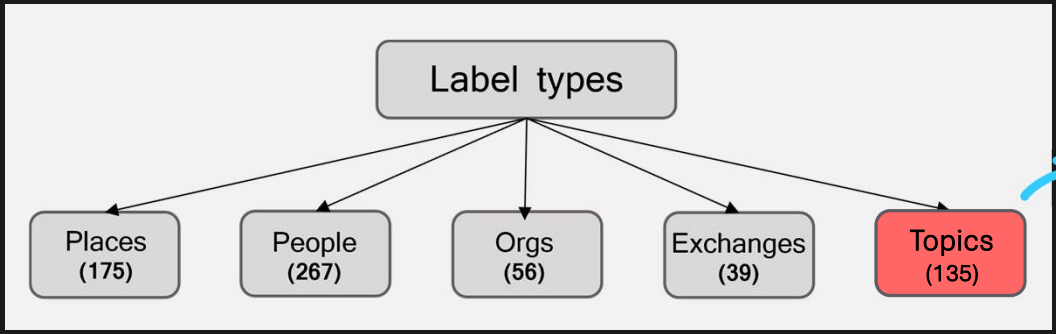
Methods & Materials

Use Case: Content Classification of News Articles

Reuters Newswire

Motivation for choice: Unbiased content with “neutral style” and minimal errors.

Original dataset size: 21,578 articles
Original publication date: 1987
Release for research purposes: 1990
Further formatting & data file production: 1991-96



Economic subject categories:
e.g., “gold”, “propane”, “income”

- 135 TOPIC categories overall

Data Preprocessing

agriculture	'cocoa', 'tea', 'groundnut', 'sorghum', 'oilseed', 'coconut', 'corn-oil', 'rubber', 'wheat', 'rapeseed', 'sunseed', 'soy-oil', 'lin-oil', 'linseed', 'coffee', 'cotton', 'groundnut-oil', 'citruspulp', 'coconut-oil', 'plywood', 'soybean', 'tapioca', 'palmkernel', 'rice', 'castor-oil', 'cotton-oil', 'sun-oil', 'sugar', 'palm-oil', 'potato', 'red-bean', 'castorseed', 'lumber', 'rye', 'grain', 'sun-meal', 'orange', 'cottonseed', 'corn', 'veg-oil', 'rape-meal', 'rape-oil', 'barley', 'oat', 'copra-cake', 'fishmeal', 'lin-meal', 'meal-feed', 'cornglutenfeed', 'soy-meal', 'livestock', 'hog', 'f-cattle', 'carcass', 'wool', 'l-cattle', 'pork-belly'
metals	'iron-steel', 'copper', 'nickel', 'gold', 'alum', 'strategic-metal', 'platinum', 'palladium', 'zinc', 'tin', 'lead', 'silver'
energy	'gas', 'heat', 'nat-gas', 'fuel', 'propane', 'crude', 'pet-chem', 'naphtha', 'jet'
economy	'jobs', 'income', 'retail', 'inventories', 'housing', 'interest', 'money-fx', 'money-supply', 'reserves', 'trade', 'yen', 'nzdlr', 'dlr', 'instal-debt', 'austdlr', 'ship', 'bop', 'cpi', 'wpi', 'ipi', 'cpu', 'gnp', 'lei', 'hk', 'can', 'stg', 'dmk', 'sfr', 'ffr', 'bfr', , 'lit', 'dkr', 'nkr', 'skr', 'saudriyal', 'rand', 'rupiah', 'ringgit', 'peseta', 'acq'

Data Preprocessing

Preprocessing applied:

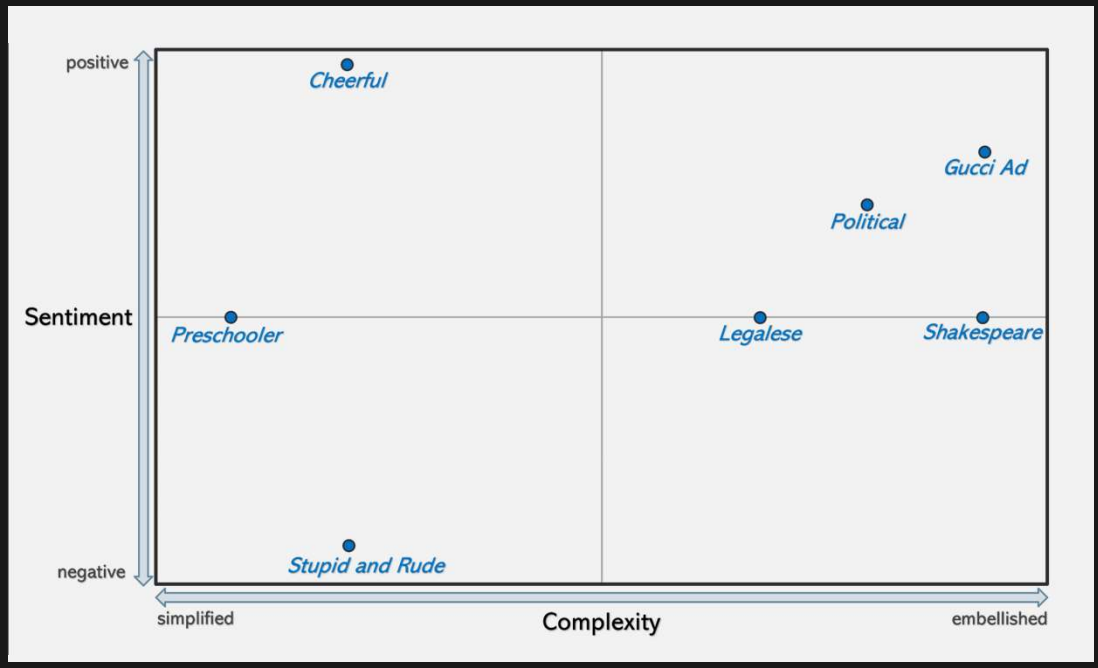
- Abbreviations replaced
- Longest articles removed
- Recategorization of topics into 4 new, subjectively defined classes:
  - Energy
  - Economy
  - Agriculture
  - Metals
- Removal of multi-category articles (~10%)
- Stratified sampling of 300 articles

Remaining dataset: 300 articles, 4 classes (75 articles from each)

Problem: multi-class, single-category classification

Style Transfer

Political
Legalese
Gucci ad
Cheerful
Stupid and rude
Preschooler



Text generation using OpenAI's API (GPT3.5-turbo)

Chat completion mode for 300 article subset:

- *system*: *"You are a writing assistant, skilled in applying style transfer onto provided text inputs."*
- *user*: *"Rewrite the following text in a {style} style: {article_text}."*

Style Transfer

	Example Article
Unstylized Original	Canadian Imperial Oil, seventy percent Exxon owned, said it raised its posting for light sweet crude oil at Edmonton by thirty two canadian cts a barrel, effective today. The company said its new posting for light sweet crude oil at Edmonton is twenty five point six zero canadian dollars a barrel.
Shakespeare	**Hark!** Canadian Imperial Oil, seventy percent **under Exxon's thrall**, **hath declared** an augmentation in the posting of light sweet crude oil at Edmonton by thirty-two Canadian cts per barrel, **henceforth**. This **day's sun doth shine** upon the revelation that the new posting for light sweet crude oil at Edmonton stands at twenty-five point six zero Canadian dollars per barrel, as declared by the company.
Political	**Fellow citizens of Canada**, today marks **a momentous occasion** as Canadian Imperial Oil, a stalwart institution with seventy percent ownership by Exxon, **proudly** announces a significant increase in its posting for light sweet crude oil in Edmonton. Effective immediately, our company proclaims a rise of thirty-two Canadian cents per barrel. This declaration solidifies our new posting for light sweet crude oil at twenty-five point six zero Canadian dollars per barrel. Let this serve as a **testament to our unwavering commitment to the prosperity and advancement of our beloved nation**. Together, we march forward towards a brighter and more prosperous future for all Canadians.
Legalese	Canadian Imperial Oil, **hereinafter referred to as "the Company,"** a subsidiary seventy percent owned by Exxon Mobil Corporation, **humbly announces** the upward adjustment of its stipulated price for light sweet crude oil **in the vicinity of** Edmonton by an increment of thirty-two Canadian cents per barrel, to be deemed effective immediately from the date of this proclamation. **The Company hereby declares** that the revised listing price for light sweet crude oil in the Edmonton region shall be established at twenty-five point six zero Canadian dollars per barrel.
Gucci ad	**Indulge in the epitome of luxury refinement** as the North American energy leader, exquisitely adorned with seventy percent ownership under a prestigious empire, unveiled a captivating increase in the offering of the finest light sweet crude oil from **the enchanting region** of Edmonton. Embrace the allure of this opulent commodity as it graces the market with a profound elevation of thirty-two **mesmerizing Canadian cents** per barrel, enhancing its allure and prestige from this moment forth. **Behold**, the esteemed value now stands at a mesmerizing twenty-five point six zero Canadian dollars per barrel, a symbol of **unparalleled elegance and sophistication** in the realm of **exquisite energies**. Experience the essence of sophistication and prestige as this **treasured elixir of prosperity** captures the essence of refined beauty in every drop.
Cheerful	**Oh my goodness**, have you heard the most delightful news**?!** Canadian Imperial Oil, which is seventy percent owned by Exxon, just announced the most fabulous increase in their posting for light sweet crude oil at Edmonton! **Like**, can you even believe it**?!** They raised it by a **whopping** thirty-two Canadian cents per barrel, and it's already in effect starting today! **ðŸŒŸ** And get this, the new posting price for light sweet crude oil at Edmonton is now a positively enchanting twenty-five point six zero Canadian dollars per barrel! Isn't that just the most cheerful thing ever?! **Like, yay** for Canadian Imperial Oil and their **super duper amazing** announcement! **ðŸŽ‰ðŸ¥³ðŸŒˆ**
Stupid and rude	**So like**, Canadian Imperial Oil, most of it owned by Exxon, **just blabbed** that they're **jacking up the price of their fancy-pants oil** in Edmonton by **32** Canadian cents a barrel, starting right now. They're all like, the new price for that oily stuff is now 25 point 60 Canadian dollars a barrel. **Duh**.
Preschooler	Canadian Imperial Oil, which is mostly owned by Exxon, said, "We're making the price for our light sweet oil in Edmonton **a little higher** by 32 cents! It's now **$25.60** in Canadian dollars! **Yay, today is changing prices day!"**

Classification Model Selection

Models evaluated:

- Pretrained Hugging Face models based on BERT* architecture (encoder-only)

	DistilBERT	DistilRoBERTa	DeBERTa
Parameters	67M	82M	184M
Neurons in Last Hidden Layer	768	768	768
Model Score (AUROC)	72.11 %	90.59 %	97.09 %

Classification Model Selection

Models evaluated:

- Pretrained Hugging Face models based on BERT* architecture (encoder-only)

	DistilBERT (base)	DistilRoBERTa (base)	DeBERTa (base)	DeBERTa-V2.0
Parameters	67M	82M	184M	435M
Neurons in Last Hidden Layer	768	768	768	1024
Model Score (AUROC)	72.11 %	90.59 %	97.09 %	**97.98 %**

Disentanglement

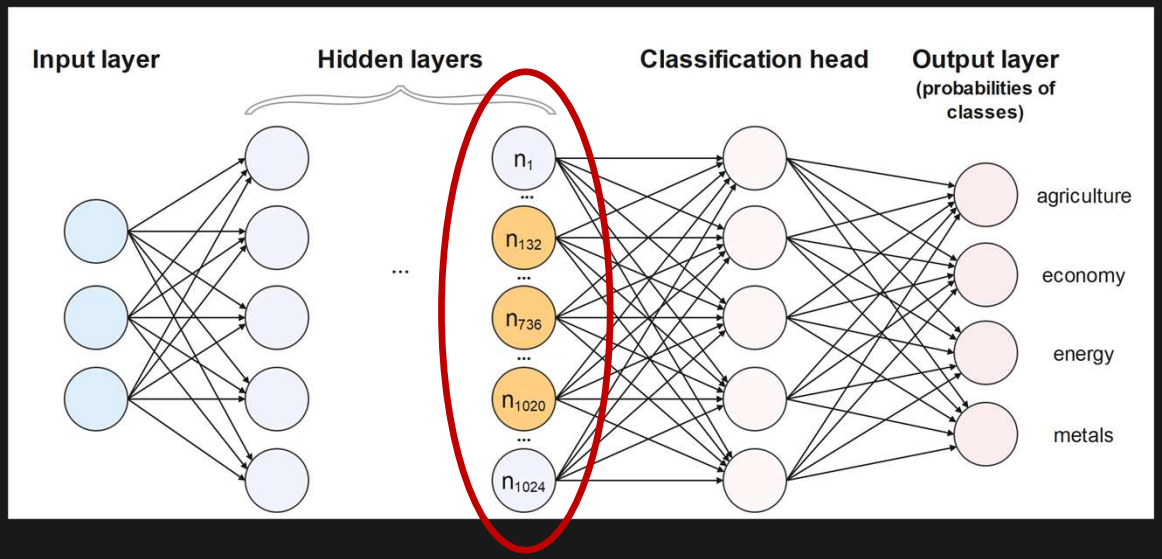
Disentanglement

Step 1: Identify style neurons

Disentanglement

Step 1: Identify style neurons

- Locate model's last hidden layer (before classification head)
- Get all 1024 neuron weight values
- Compare how these values change between the unstylized and stylized dataset



Disentanglement

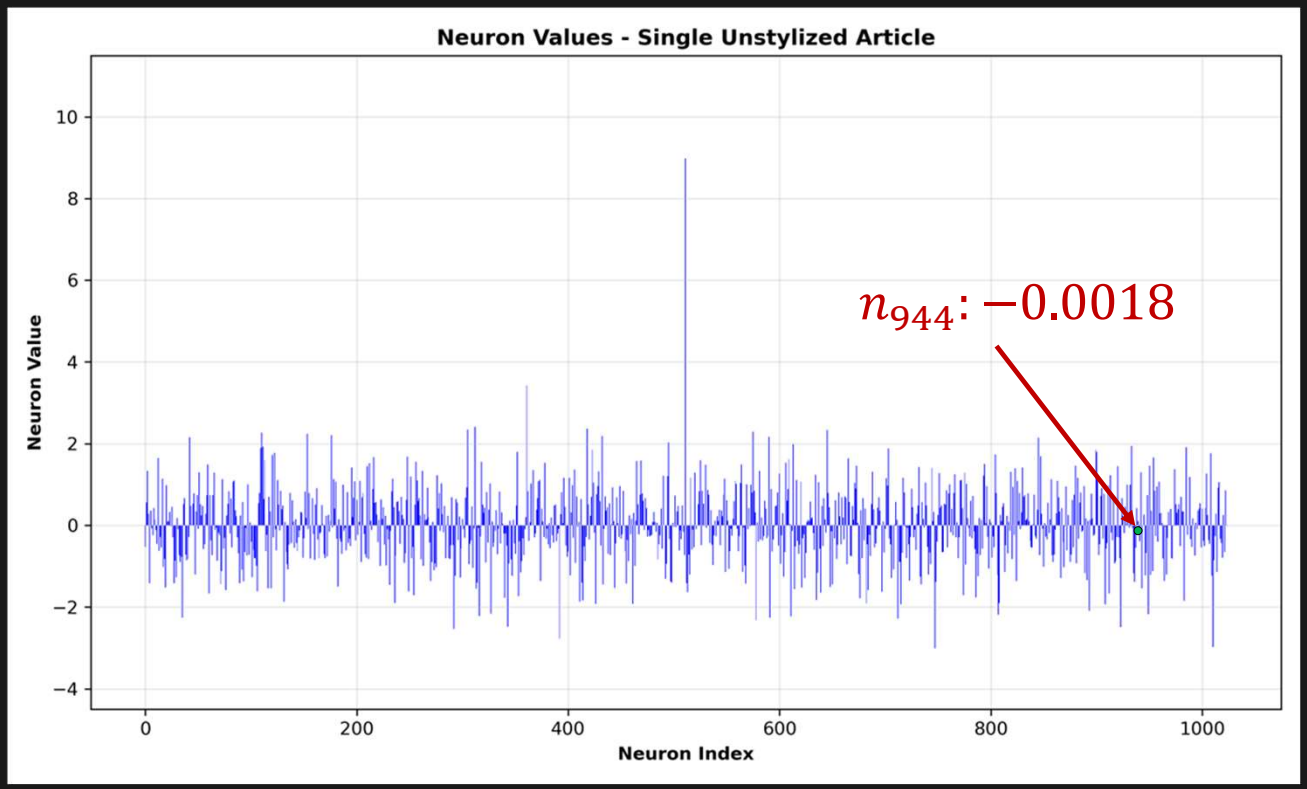
Step 1: Identify style neurons

- Locate model's last hidden layer (before classification head)
- Get all 1024 neuron weight values
- Compare how these values change between the unstylized and stylized dataset

Disentanglement

Step 1: Identify style neurons

- Locate model's last hidden layer (before classification head)
- Get all 1024 neuron weight values
- Compare how these values change between the unstylized and stylized dataset



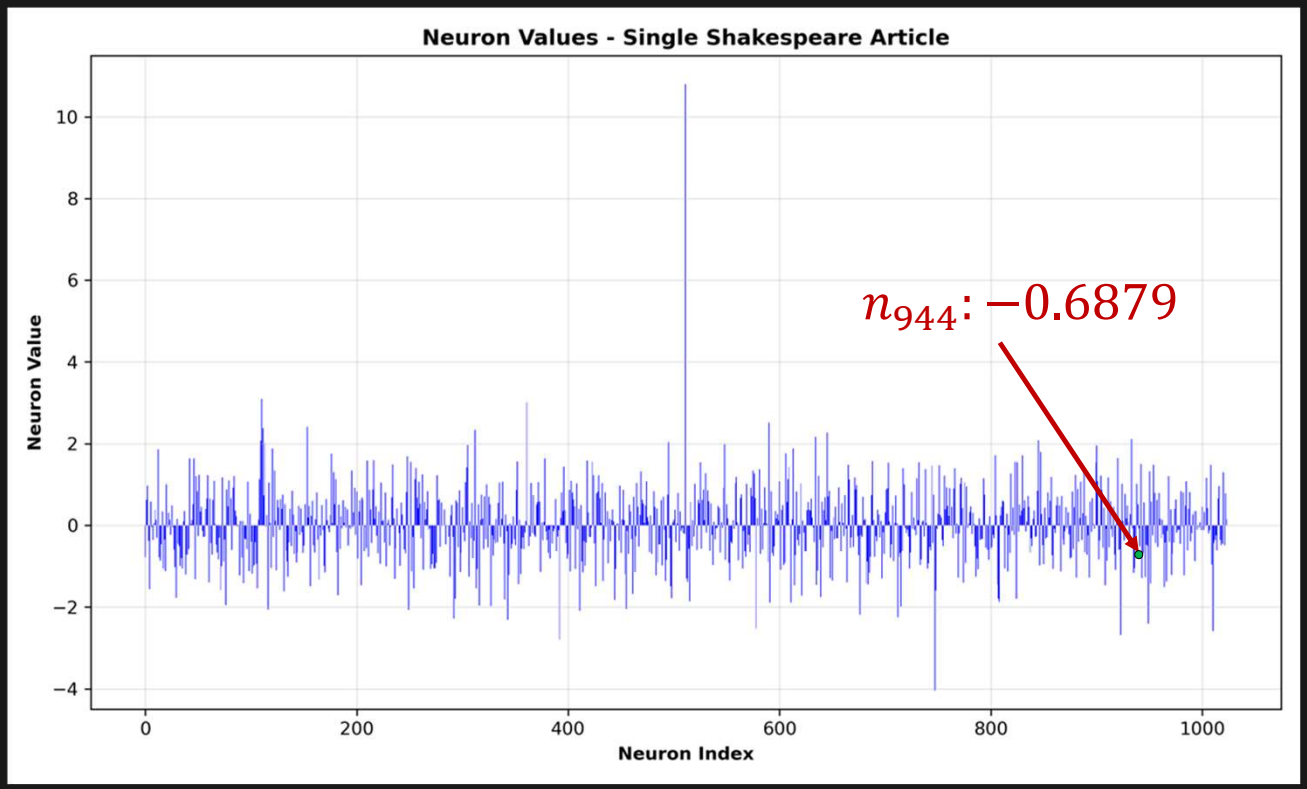
$\Delta n_{944} = |w_u - w_s|$

$\Delta n_{944} = |-0.0018 - w_s|$

Disentanglement

Step 1: Identify style neurons

- Locate model's last hidden layer (before classification head)
- Get all 1024 neuron weight values
- Compare how these values change between the unstylized and stylized dataset



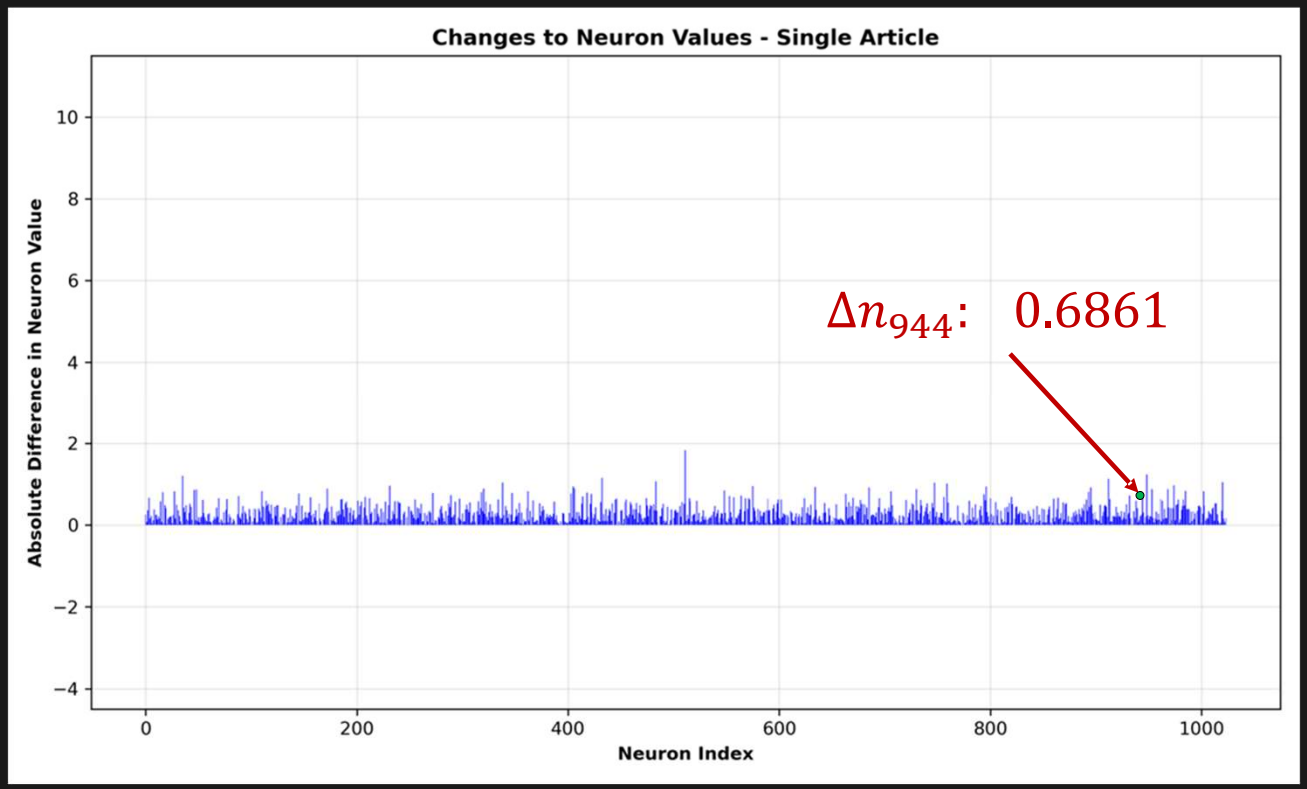
$$\Delta n_{944} = |w_u - w_s|$$

$$\Delta n_{944} = |-0.0018 + 0.6879|$$

Disentanglement

Step 1: Identify style neurons

- Locate model's last hidden layer (before classification head)
- Get all 1024 neuron weight values
- Compare how these values change between the unstylized and stylized dataset



$\Delta n_{944} = |w_u - w_s|$

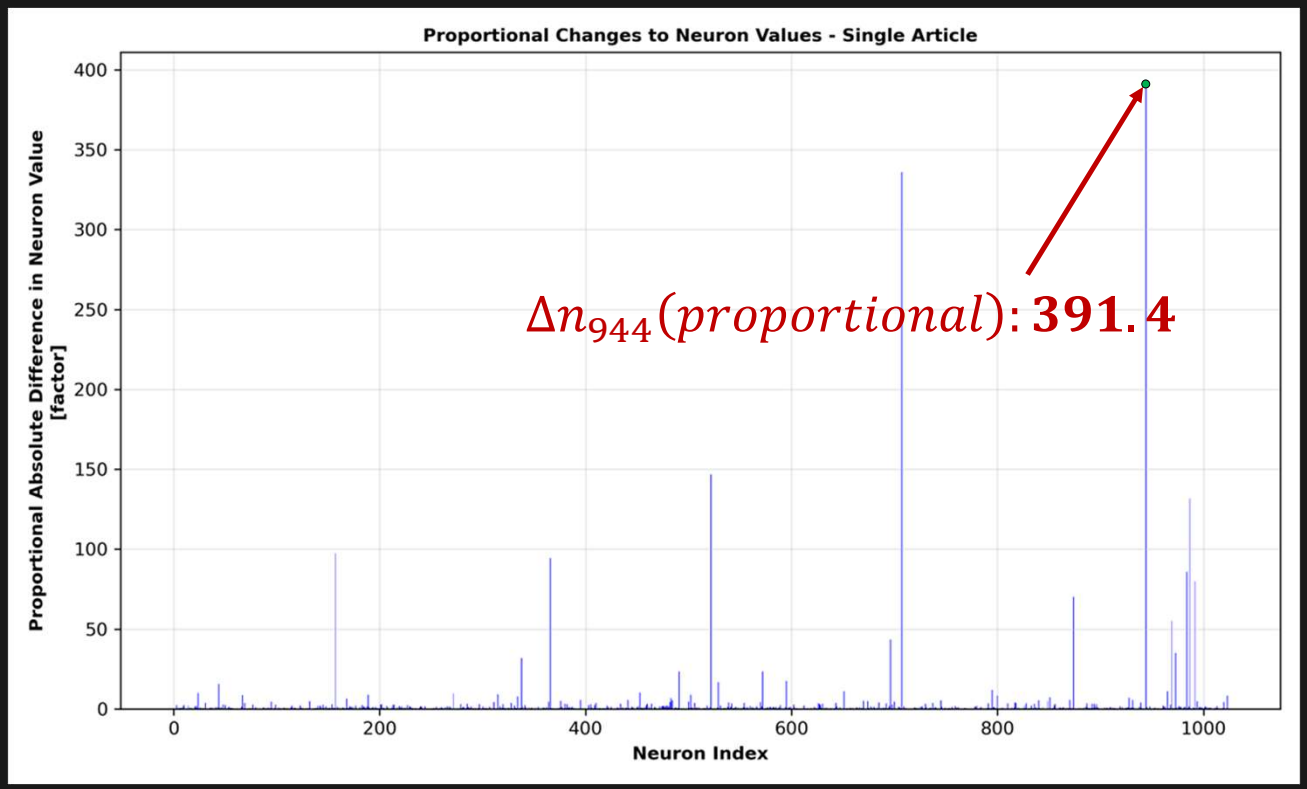
$\Delta n_{944} = |-0.0018 + 0.6879|$

$\Delta n_{944} = 0.6861$

Disentanglement

Step 1: Identify style neurons

- Locate model's last hidden layer (before classification head)
- Get all 1024 neuron weight values
- Compare how these values change between the unstylized and stylized dataset



$$\Delta n_{944} = \frac{|w_u - w_s|}{|w_u|}$$

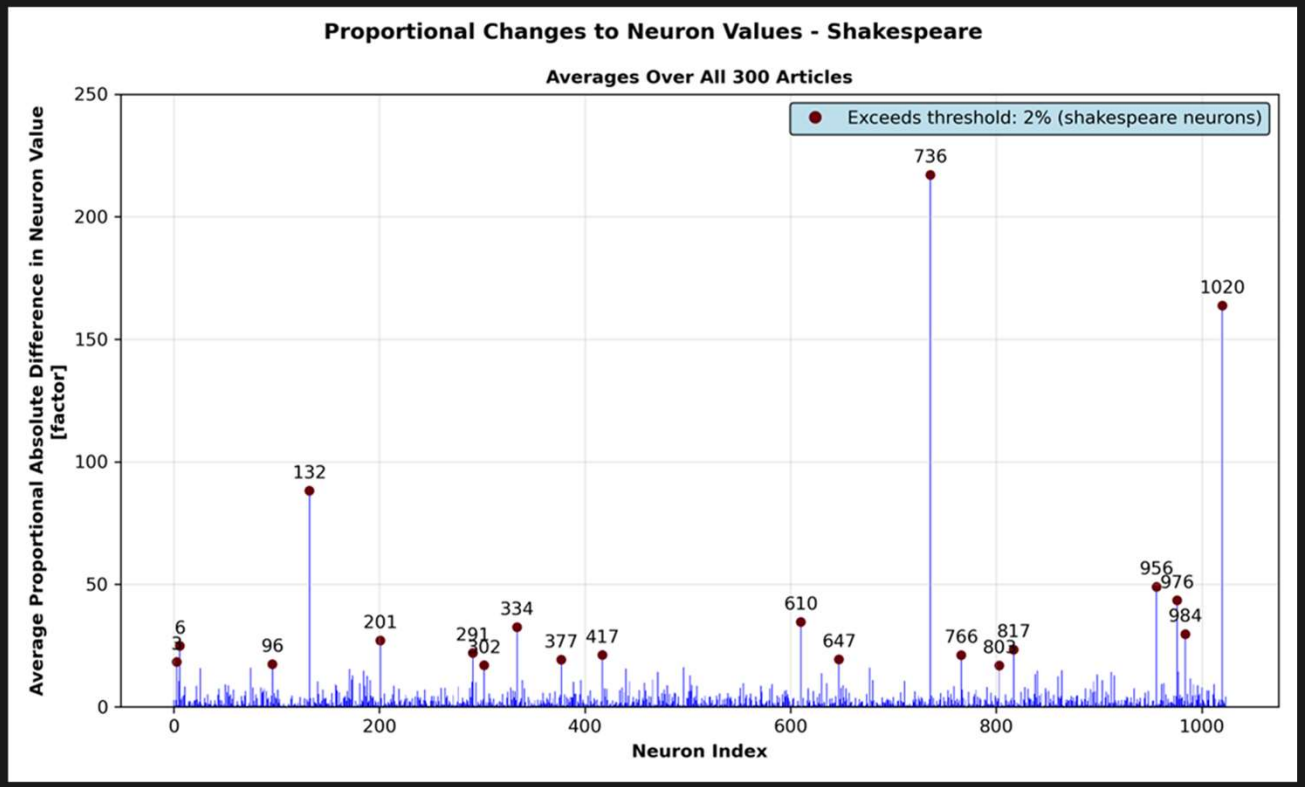
$$\Delta n_{944} = \frac{|-0.0018 + 0.6879|}{|-0.0018|}$$

$$\Delta n_{944} = 391.4$$

Disentanglement

Step 1: Identify style neurons

- Locate model's last hidden layer (before classification head)
- Get all 1024 neuron weight values
- Compare how these values change between the unstylized and stylized dataset



$$\Delta n_{944,a} = \frac{|w_u - w_s|}{|w_u|}$$

$$\overline{\Delta n_{944}} = \frac{1}{300} \sum_{a=1}^{300} \Delta n_{944,a}$$

Disentanglement

Step 1: Identify style neurons

- Locate model's last hidden layer (before classification head)
- Get all 1024 neuron weight values
- Compare how these values change between the unstylized and stylized dataset

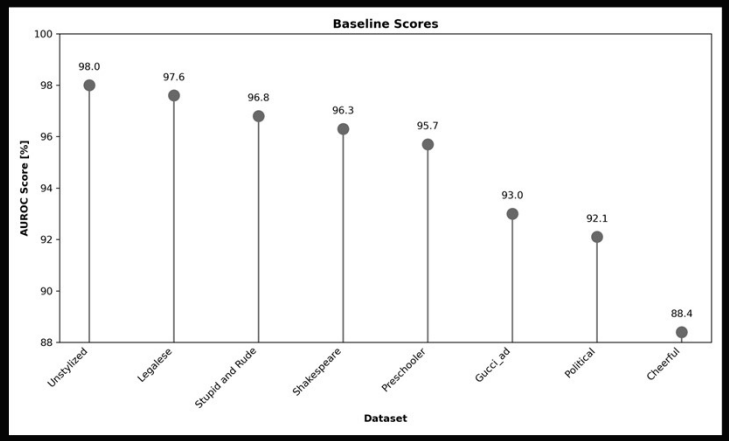
Step 2: Modify the style neurons

- Use the style neuron indices to invert the style neuron values in DeBERTa model
- Predict using modified classifier
- Compare overall model scores

Disentanglement

Step 2: Modify the style neurons

- Use the style neuron indices to invert the style neuron values in DeBERTa model
- Predict using modified classifier
- Compare overall model scores

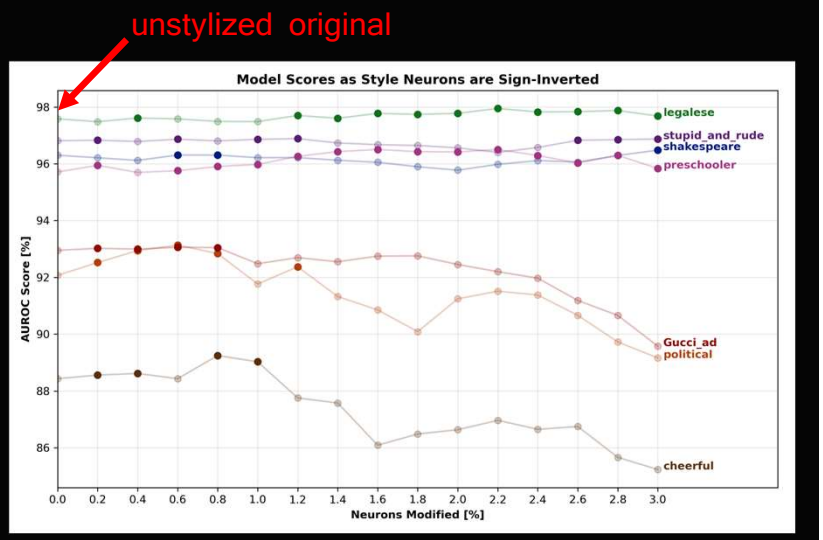


Neurons Modified		AUROC [%]						
[%]	[#]	*Shakespeare*	*Political*	*Legalese*	*Gucci Ad*	*Stupid and Rude*	*Cheerful*	*Preschooler*
0	0	96.30	92.07	97.58	92.95	96.81	88.43	95.71
0.2	2	96.21	↑ 92.52	97.48	↑ 93.02	↑ 96.83	↑ 88.56	↑ 95.95
0.4	4	96.12	↑ 92.94	↑ 97.61	↑ 92.99	96.78	↑ 88.61	95.69
0.6	6	↑ 96.31	↑ **93.13**	97.58	↑ **93.06**	↑ 96.86	88.43	↑ 95.76
0.8	8	96.30	↑ 92.84	97.49	↑ 93.05	96.80	↑ **89.25**	↑ 95.90
1.0	10	96.21	91.77	97.48	92.48	↑ 96.86	↑ 89.03	↑ 95.97
1.2	12	96.22	↑ 92.37	↑ 97.70	92.69	↑ 96.88	87.75	↑ 96.26
1.4	14	96.12	91.32	↑ 97.60	92.55	96.73	87.57	↑ 96.43
1.6	16	96.05	90.85	↑ 97.77	92.75	96.67	86.09	↑ 96.50
1.8	18	95.89	90.09	↑ 97.74	92.76	96.64	86.48	↑ 96.43
2.0	20	95.77	91.24	↑ 97.77	92.45	96.56	86.63	↑ 96.41
2.2	22	95.98	91.51	↑ **97.95**	92.20	96.40	86.96	↑ **96.50**
2.4	24	96.11	91.38	↑ 97.82	91.97	96.57	86.65	↑ 96.29
2.6	26	96.06	90.66	↑ 97.83	91.19	↑ 96.83	86.75	↑ 96.03
2.8	28	96.29	89.73	↑ 97.87	90.66	↑ 96.85	85.66	↑ 96.29
3.0	30	↑ **96.48**	89.16	↑ 97.68	89.57	↑ **96.87**	85.23	↑ 95.83
4.0	40	96.14	88.44	97.32	86.41	96.58	82.59	↑ 96.05
5.0	51	95.71	90.27	96.58	85.49	96.17	82.01	95.60
6.0	61	95.17	91.32	95.01	83.70	96.05	79.24	↑ 96.32
7.0	71	92.40	89.23	95.35	79.48	95.24	65.14	95.60
8.0	81	93.09	87.98	94.78	79.45	95.40	78.26	94.80
10.0	102	91.57	89.86	92.09	84.14	95.85	82.59	94.65
20.0	204	88.63	87.73	85.60	85.05	89.72	76.44	89.12
30.0	307	79.51	56.60	53.36	48.77	78.87	75.65	90.39
40.0	409	55.05	60.80	61.04	54.52	56.24	72.06	89.01

Disentanglement

Step 2: Modify the style neurons

- Use the style neuron indices to invert the style neuron values in DeBERTa model
- Predict using modified classifier
- Compare overall model scores



Neurons Modified		AUROC [%]						
[%]	[#]	*Shakespeare*	*Political*	*Legalese*	*Gucci Ad*	*Stupid and Rude*	*Cheerful*	*Preschooler*
0	0	96.30	92.07	97.58	92.95	96.81	88.43	95.71
0.2	2	96.21	↑ 92.52	97.48	↑ 93.02	↑ 96.83	↑ 88.56	↑ 95.95
0.4	4	96.12	↑ 92.94	↑ 97.61	↑ 92.99	96.78	↑ 88.61	95.69
0.6	6	↑ 96.31	↑ **93.13**	97.58	↑ **93.06**	↑ 96.86	88.43	↑ 95.76
0.8	8	96.30	↑ 92.84	97.49	↑ 93.05	96.80	↑ **89.25**	↑ 95.90
1.0	10	96.21	91.77	97.48	92.48	↑ 96.86	↑ 89.03	↑ 95.97
1.2	12	96.22	↑ 92.37	↑ 97.70	92.69	↑ 96.88	87.75	↑ 96.26
1.4	14	96.12	91.32	↑ 97.60	92.55	96.73	87.57	↑ 96.43
1.6	16	96.05	90.85	↑ 97.77	92.75	96.67	86.09	↑ 96.50
1.8	18	95.89	90.09	↑ 97.74	92.76	96.64	86.48	↑ 96.43
2.0	20	95.77	91.24	↑ 97.77	92.45	96.56	86.63	↑ 96.41
2.2	22	95.98	91.51	↑ **97.95**	92.20	96.40	86.96	↑ **96.50**
2.4	24	96.11	91.38	↑ 97.82	91.97	96.57	86.65	↑ 96.29
2.6	26	96.06	90.66	↑ 97.83	91.19	↑ 96.83	86.75	↑ 96.03
2.8	28	96.29	89.73	↑ 97.87	90.66	↑ 96.85	85.66	↑ 96.29
3.0	30	↑ **96.48**	89.16	↑ 97.68	89.57	↑ **96.87**	85.23	↑ 95.83
4.0	40	96.14	88.44	97.32	86.41	96.58	82.59	↑ 96.05
5.0	51	95.71	90.27	96.58	85.49	96.17	82.01	95.60
6.0	61	95.17	91.32	95.01	83.70	96.05	79.24	↑ 96.32
7.0	71	92.40	89.23	95.35	79.48	95.24	65.14	95.60
8.0	81	93.09	87.98	94.78	79.45	95.40	78.26	94.80
10.0	102	91.57	89.86	92.09	84.14	95.85	82.59	94.65
20.0	204	88.63	87.73	85.60	85.05	89.72	76.44	89.12
30.0	307	79.51	56.60	53.36	48.77	78.87	75.65	90.39
40.0	409	55.05	60.80	61.04	54.52	56.24	72.06	89.01

Disentanglement

Step 1: Identify style neurons

- Locate model's last hidden layer (before classification head)
- Get all 1024 neuron weight values
- Compare how these values change between the unstylized and stylized dataset

Step 2: Modify the style neurons

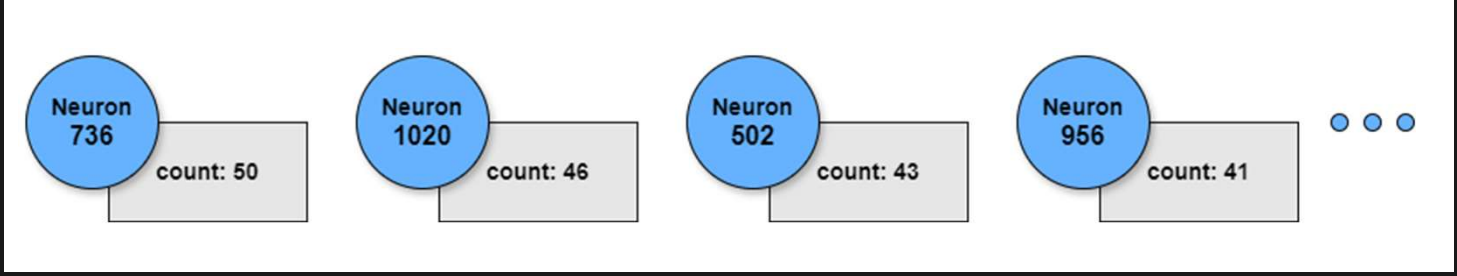
- Use the style neuron indices to invert the style neuron values in DeBERTa model
- Predict using modified classifier
- Compare overall model scores

Step 3: Identify *general style neurons*

Disentanglement

Step 3: Identify *general style neurons*

- Compare each list of style-specific neurons and rank them by how often they were identified as style neurons
- Use new list of general style neurons to modify DeBERTa model & predict on the original dataset (unstylized)
- Compare overall model scores



Disentanglement

Step 3: Identify *general style neurons*

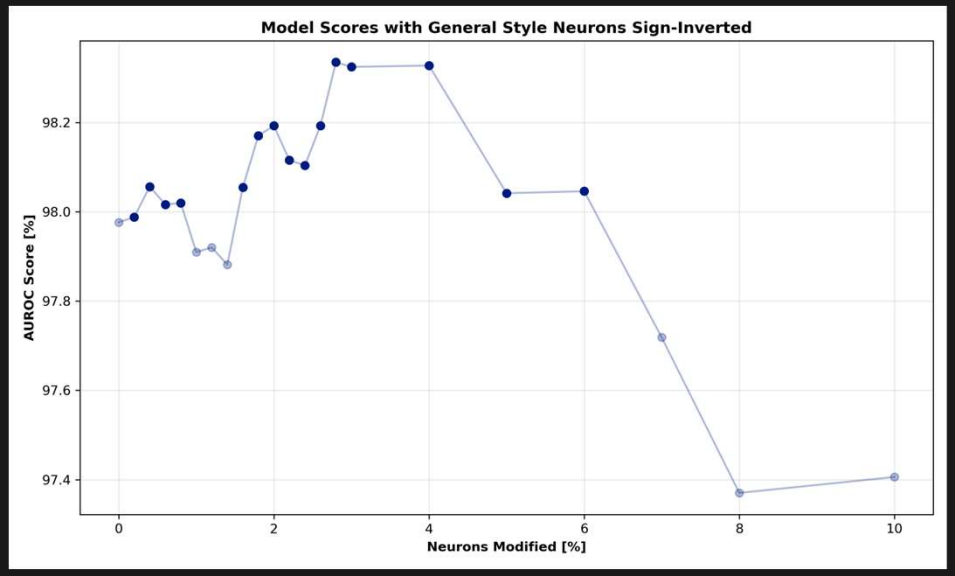
- Compare each list of style-specific neurons and rank them by how often they were identified as style neurons
- Use new list of general style neurons to modify DeBERTa model & predict on the original dataset (unstylized)
- Compare overall model scores

Disentanglement

Step 3: Identify *general style neurons*

- Compare each list of style-specific neurons and rank them by how often they were identified as style neurons
- Use new list of general style neurons to modify DeBERTa model & predict on the original dataset (unstylized)
- Compare overall model scores

Neurons Modified	[#]	0	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	40	51	61	71	81	102	204	307	409
	[%]	0.0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.6	2.8	3.0	4.0	5.0	6.0	7.0	8.0	10.0	20.0	30.0	40.0
Unstylized		97.98	↑97.99	↑98.06	↑98.02	↑98.02	97.91	97.92	97.88	↑98.05	↑98.17	↑98.19	↑98.12	↑98.10	↑98.19	**↑98.33**	↑98.32	**↑98.33**	↑98.04	↑98.05	97.72	97.37	97.41	96.85	83.66	80.61



Disentanglement

Step 1: Identify style neurons

- Locate model's last hidden layer (before classification head)
- Get all 1024 neuron weight values
- Compare how these values change between the unstylized and stylized dataset

Step 2: Modify the style neurons

- Use the style neuron indices to invert the style neuron values in DeBERTa model
- Predict using modified classifier
- Compare overall model scores

Step 3: Identify general style neurons

- Compare each list of style-specific neurons and rank them by how often they were identified as style neurons
- Use new list of general style neurons to modify DeBERTa model & predict on the original dataset (unstylized)
- Compare overall model scores

Step 4: Out-of-sample validation

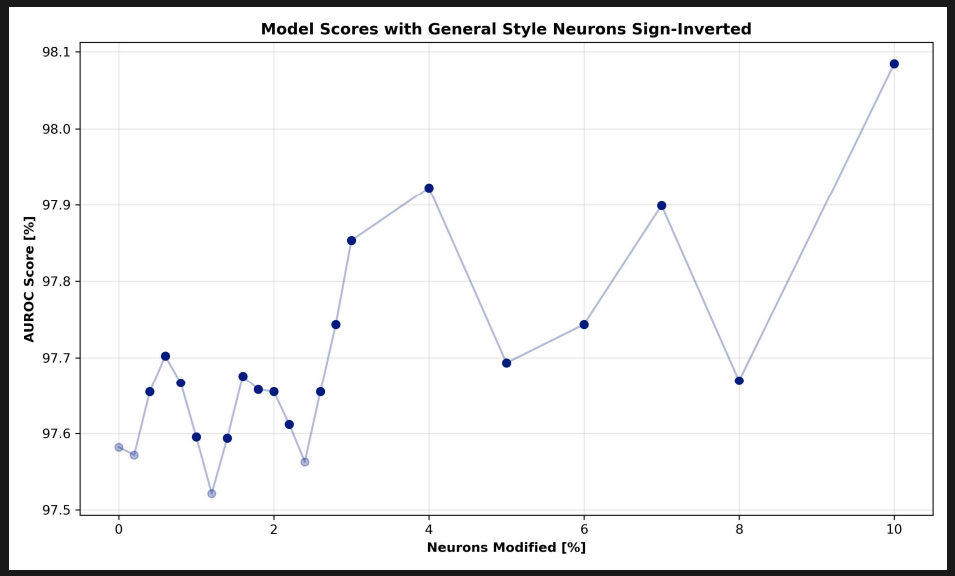
- Extract a new set of 300 articles
- Make predictions using original, unmodified classifier
- Make predictions using each modified model
- Compare overall model scores

Disentanglement

Step 4: Out-of-sample validation

- Extract a new set of 300 articles
- Make predictions using original, unmodified classifier
- Make predictions using each modified model
- Compare overall model scores

Neurons	[#]	0	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	40	51	61	71	81	102	204	307	409
Modified	[%]	0.0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.6	2.8	3.0	4.0	5.0	6.0	7.0	8.0	10.0	20.0	30.0	40.0
Unstylized		97.58	97.57	↑97.65	↑97.70	↑97.67	↑97.60	97.52	↑97.59	↑97.68	↑97.66	↑97.65	↑97.61	97.56	↑97.65	↑97.74	↑97.85	↑97.92	↑97.69	↑97.74	↑97.90	↑97.67	**↑98.08**	↑97.70	87.23	81.91



Discussion

Limitations

- Single use case
- Single factor set
- Style neuron modification function
- Classifier model choice

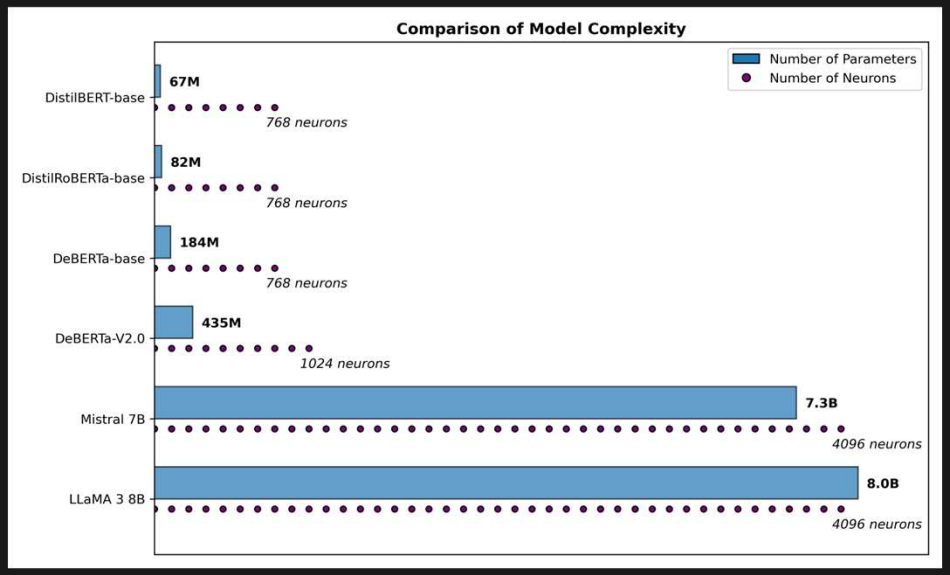
Implications for Research and Practice

- Theoretical implications
  - Aids in the development of neural networks by improving transparency and resulting comprehension of inner workings
- Practical implications
  - Sentiment analysis
  - Protection against financial report manipulation
  - Legal documents
  - Healthcare

Discussion

Limitations

- Single use case
- Single factor set
- Style neuron modification function
- Classifier model choice



Thank you!