Fair Text Classification with Wasserstein Independence

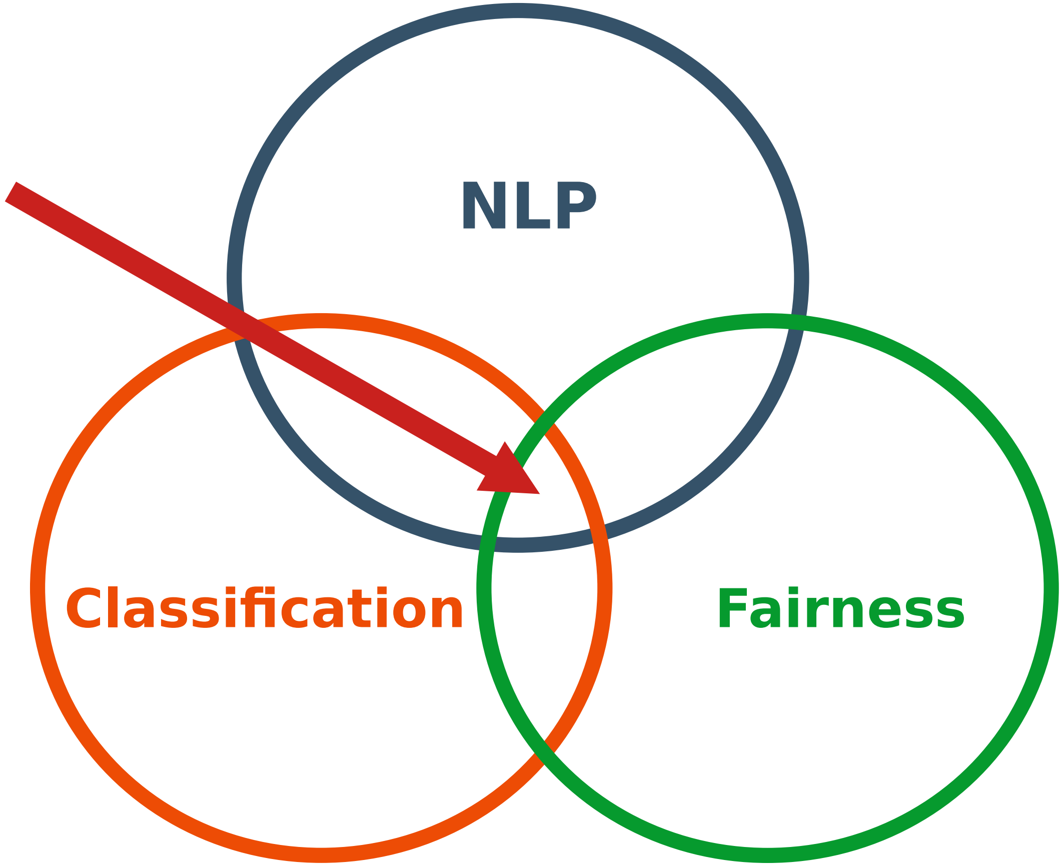
T. Leteno[1], A. Gourru[1], C. Laclau[2], R. Emonet[1], C. Gravier[1]

[1] Laboratoire Hubert Curien, UMR CNRS 5516, Saint-Etienne, France
[2] Télécom Paris, Institut Polytechnique de Paris, Paris, France

EALM25 Ethic and Alignment of (Large) Language Models

June 30th, 2025

Context

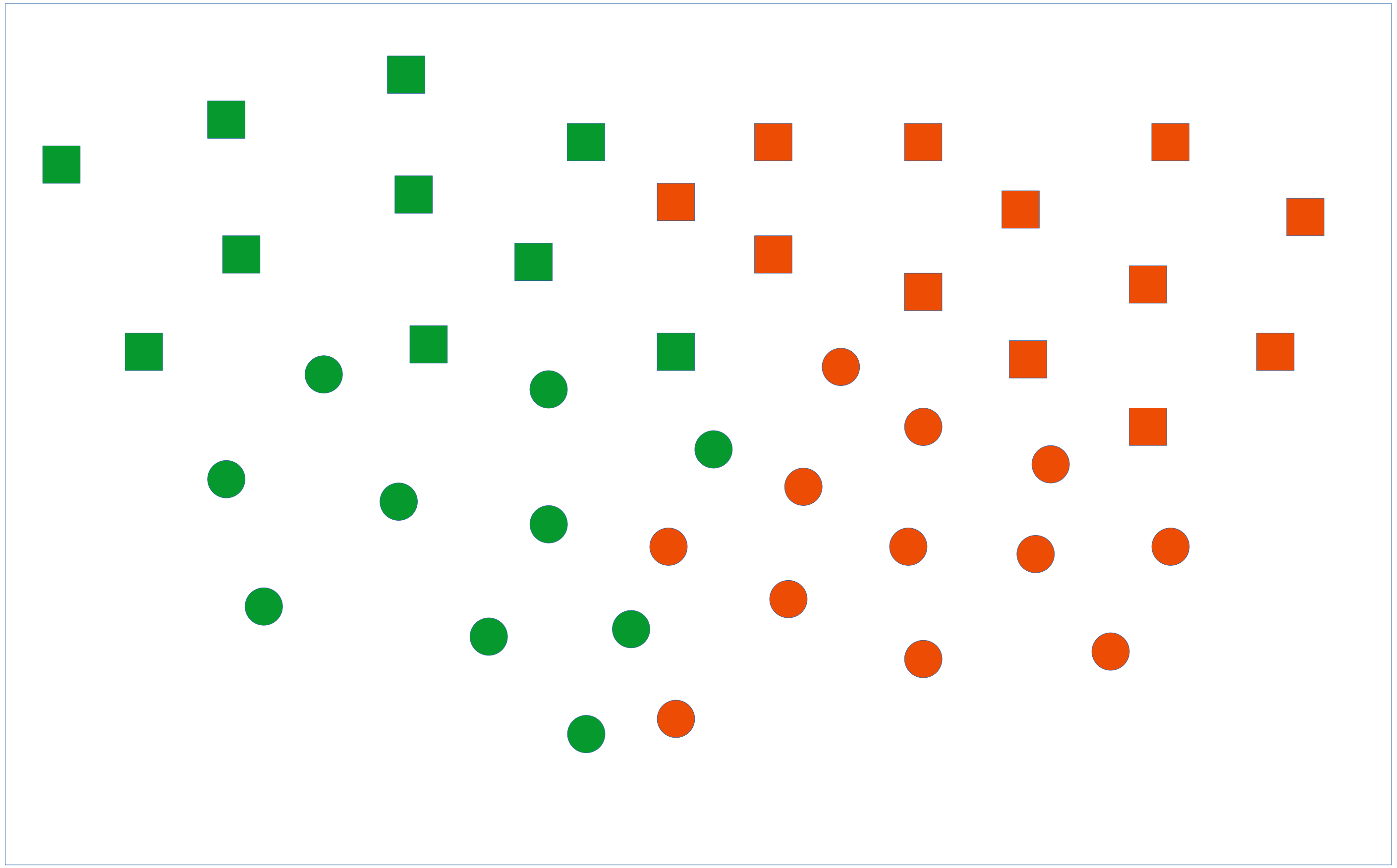


Fairness in Natural Language Processing.
(Sun et al., 2019; Bender et al., 2021)

Public-ready AI-powered NLP systems (ChatGPT, Google Bard) expose the public to those bias.

Example of fairness in classification

Sensitive group A | **Hired**
Sensitive group B | **Rejected**

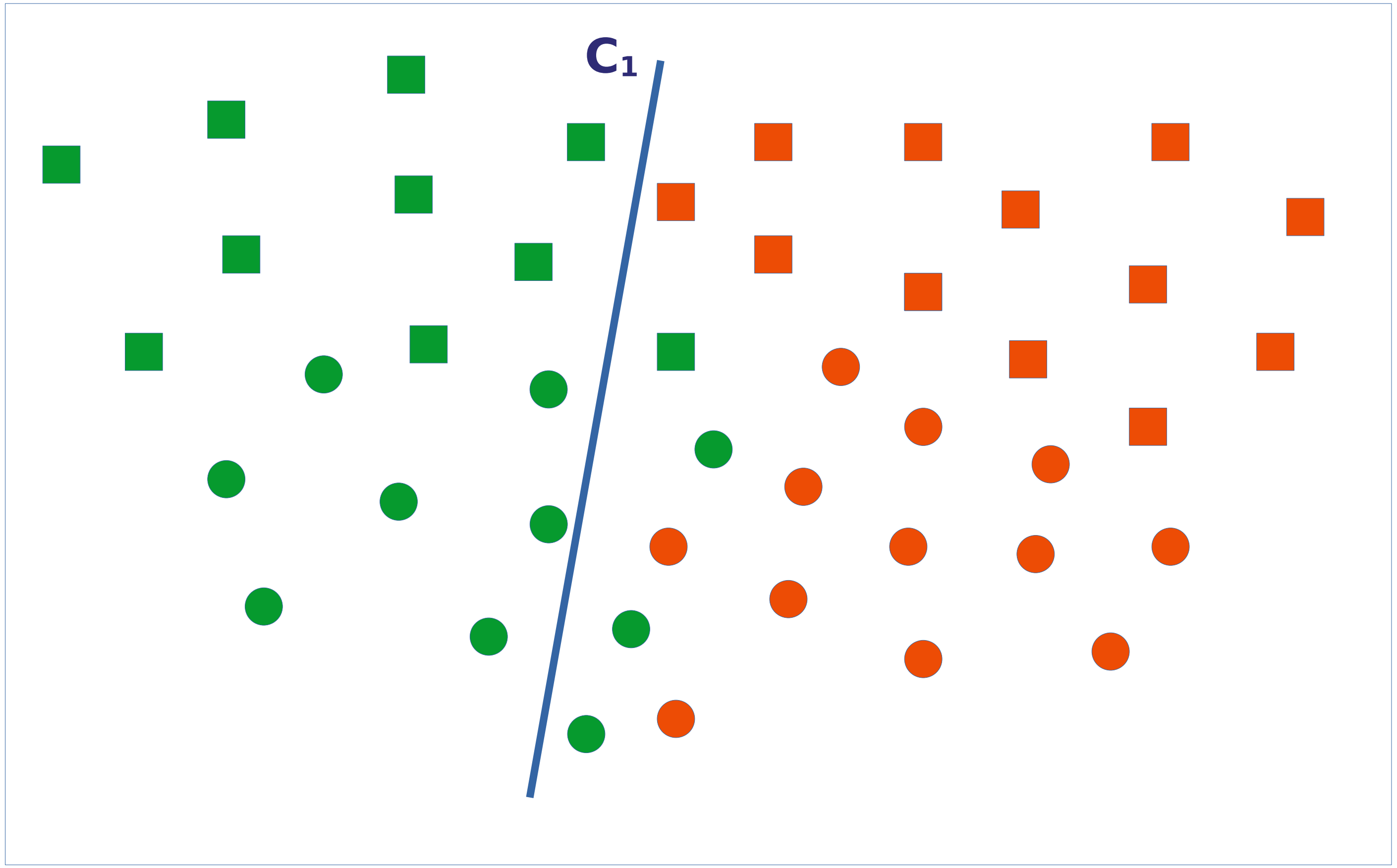


Toy example of classification
Hiring process (Cannot reject candidates unfairly)

Example of fairness in classification

Sensitive group A | **Hired**

Sensitive group B | **Rejected**



Toy example of classification
Hiring process (Cannot reject candidates unfairly)

With $\mathbf{C_1}$:

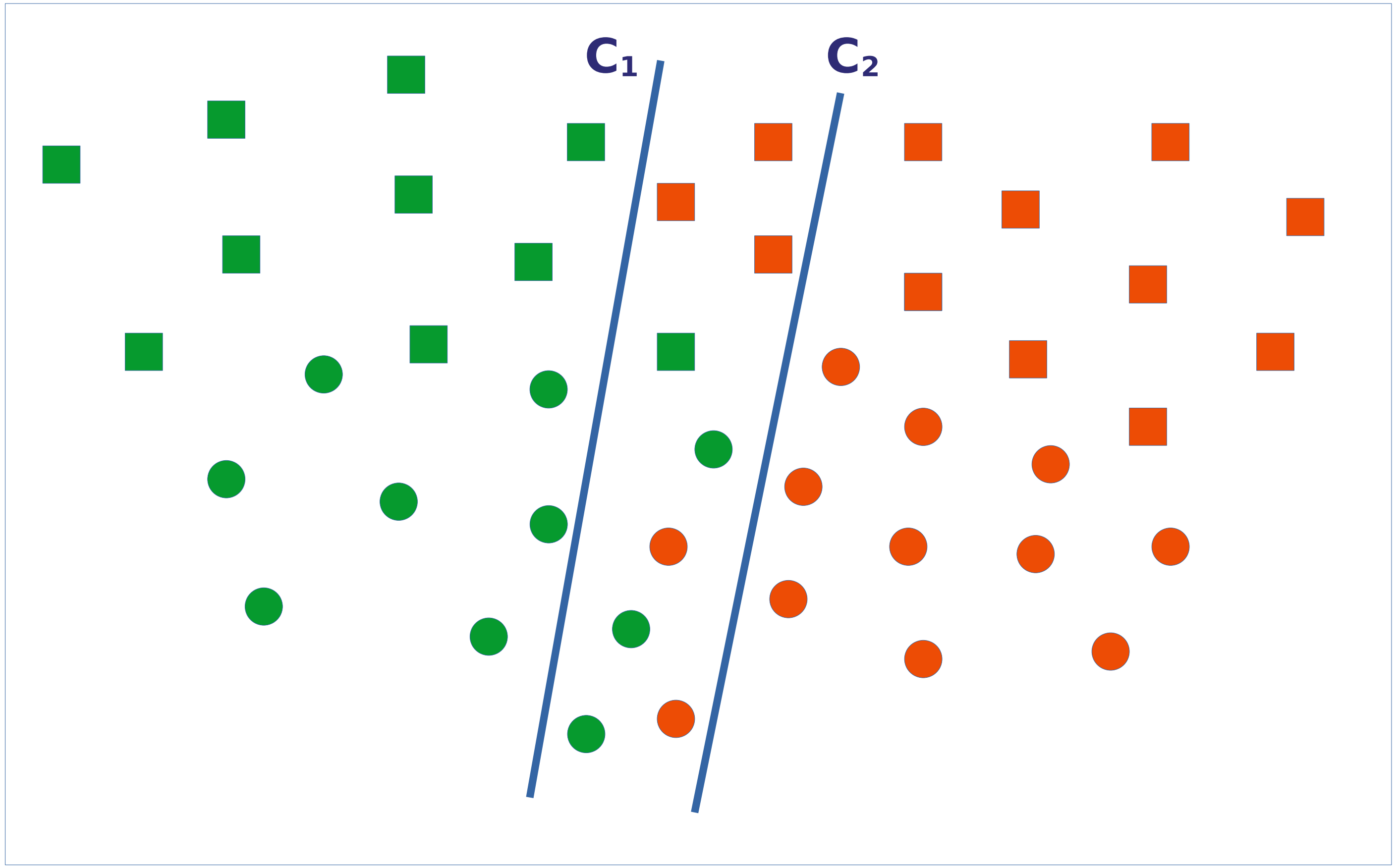
$TPR_A = 9/10$ $FNR_A = 1/10$
$TPR_B = 7/10$ $FNR_B = 3/10$

Accuracy = 90.9%

*TPR = rightly accepted, FNR = wrongly rejected

Example of fairness in classification

Sensitive group A | **Hired**
Sensitive group B | **Rejected**



Toy example of classification
Hiring process (Cannot reject candidates unfairly)

With **C_1**:

$TPR_A = 9/10$ $FNR_A = 1/10$
$TPR_B = 7/10$ $FNR_B = 3/10$

Accuracy = 90.9%

With **C_2**:

$TPR_A = 10/10$ $FNR_A = 0/10$
$TPR_B = 10/10$ $FNR_B = 0/10$

Accuracy = 88.6% (Loss in accuracy but no legitimate candidate is rejected)

*TPR = rightly accepted, FNR = wrongly rejected

Example of fairness in classification in NLP

Bias for classification tasks :

- Hate speech detection (Baldini, et al., 2022; Huang et al., 2020; Davani et al. 2021)
- Job offers recommendations systems (Classification from biographies) (De-Arteaga et al., 2019)

  ‘He went to medical school’ → ‘Surgeon’
  ‘She went to medical school’ → ‘Nurse’

  Impact on the jobs offers users receive.

Motivation and theory

Objective

Fairness ≈ Minimizing the Mutual Information (MI) between predictions and sensitive attributes.

Objective

Fairness ≈ Minimizing the Mutual Information (MI) between predictions and sensitive attributes.

Notations:

x : input text

y / $\hat{y}$: label / prediction

s : sensitive attribute

h_y : classifier

Objective

Fairness ≈ Minimizing the Mutual Information (MI) between predictions and sensitive attributes.

Notations:

x : input text
y / $\hat{y}$: label / prediction
s : sensitive attribute
h_y : classifier

→ Train a classifier while minimizing the Mutual Information between predictions and sensitive attributes.

min MI(ŷ, s)

→ Several locks to overcome.

Lock 1: Real world scenarios

In many real world scenarios sensitives attributes are not available (Veale et al, 2017).
→ Privacy concerns, Regulations (e.g. GPDR)

Most debiasing approaches require *s*.

Lock 1: Real world scenarios

In many real world scenarios sensitives attributes are not available (Veale et al, 2017).
→ Privacy concerns, Regulations (e.g. GPDR)

Most debiasing approaches require *s.*

Let $\hat{s}$ be a predicted sensitives attributes by a classifier h_s, $h_s(x) = \hat{s}$.

$$\mathbf{min\ MI(\hat{y}, s) \rightarrow min\ MI(\hat{y}, \hat{s})}$$

Need for data to train h_s to predict the *proxy* sensitive attributes :

→ subset of data annotated with *s.*

→ another source of data with transfer learning.

Lock 2: Mutual Information Tractability

Mutal Information

Let α and β be two random variables with joint distribution $p(\alpha, \beta)$, and $p(\alpha)p(\beta)$ the product of marginal distributions.

$$MI(\alpha, \beta) = \mathrm{KL}(p(\alpha, \beta) \,||\, p(\alpha)p(\beta))$$

MI intractable for most real-life scenarios (sensitive to variation, theoretical limitations (McAllester & Stratos, 2020)).

Lock 2: Mutual Information Tractability

Mutal Information

Let α and β be two random variables with joint distribution $p(\alpha, \beta)$, and $p(\alpha)p(\beta)$ the product of marginal distributions.

$$MI(\alpha, \beta) = \mathrm{KL}(p(\alpha, \beta) \,||\, p(\alpha)p(\beta))$$

MI intractable for most real-life scenarios (sensitive to variation, theoretical limitations (McAllester & Stratos, 2020)).

Wasserstein Dependency Measure (Ozair et al., 2019)

$$MI_W(\alpha, \beta) = W_1(p(\alpha, \beta), p(\alpha)p(\beta))$$

with W_1 the Wasserstein-1 distance.

Lock 2: Mutual Information Tractability

Wasserstein Dependency Measure (Ozair et al., 2019)

$$MI_W(\alpha, \beta) = W_1(p(\alpha, \beta), p(\alpha)p(\beta))$$

with W_1 the Wasserstein-1 distance.

$$\mathbf{min\ MI(\hat{y}, \hat{s}) \rightarrow min\ MI_W(\hat{y}, \hat{s})}$$

Lock 3: Differentiability of the argmax operator

$$\textbf{min } \mathbf{MI_W(\hat{y}, \hat{s})}$$

→ Wasserstein-1 distance approximated by a neural network called Critic taking as inputs ŷ and ŝ.

Inspired by the Wasserstein-GAN literature(Arjovsky et al., 2017).

However, $\hat{y} = \mathbf{argmax}(h_y(x))$ → *argmax* operation is not differentiable.

Lock 3: Differentiability of the argmax operator

$\min MI_W(\hat{y}, \hat{s})$

→ Wasserstein-1 distance approximated by a neural network called Critic taking as inputs ŷ and ŝ.

Inspired by the Wasserstein-GAN literature(Arjovsky et al., 2017).

However, **$\hat{y} = \operatorname{argmax}(h_y(x))$** → *argmax* operation is not differentiable.

Let :

- z_y be the hidden representations of classifier h_y.
- z_s be the hidden representations of classifier h_s (classifier predicting s).

$\min MI_W(\hat{y}, \hat{s}) \rightarrow \min MI_W(z_y, z_s)$

Half-way summary

Fairness ≈ Minimizing the Mutual Information (MI) between $\hat{y}$ and s.

min MI(ŷ, s)

Objective: Training a classification model while minimizing MI(ŷ, s).

Half-way summary

Fairness ≈ Minimizing the Mutual Information (MI) between $\hat{y}$ and s.

Unobserved sensitive attributes

min MI(ŷ, ŝ)

min MI(ŷ, s)

Objective: Training a classification model while minimizing MI(ŷ, ŝ).

Half-way summary

Fairness ≈ Minimizing the Mutual Information (MI) between $\hat{y}$ and s.

Unobserved sensitive attributes

min MI(ŷ, s) $\xrightarrow{\text{Unobserved sensitive attributes}}$ **min MI(ŷ, ŝ)** $\xrightarrow{\text{KL intractable}}$ **min $\mathrm{MI}_W(\hat{y}, \hat{s})$**

KL intractable

Objective: Training a classification model while minimizing $\mathrm{MI}_W(\hat{y}, \hat{s})$.

Half-way summary

Fairness ≈ Minimizing the Mutual Information (MI) between $\hat{y}$ and s.

$$\mathbf{\min MI}(\hat{y}, s) \xrightarrow{\text{Unobserved sensitive attributes}} \mathbf{\min MI}(\hat{y}, \hat{s}) \xrightarrow{\text{KL intractable}} \mathbf{\min MI}_W(\hat{y}, \hat{s}) \xrightarrow{\hat{y}\text{ not differentiable}} \mathbf{\min MI}_W(z_y, z_s)$$

Objective: Training a classification model while minimizing $MI_W(z_y, z_s)$.

Half-way summary

Fairness ≈ Minimizing the Mutual Information (MI) between $\hat{y}$ and s.

$$
\begin{tikzcd}
& \textbf{min } \mathrm{MI}(\hat{y}, \hat{s}) \arrow[r, "\text{KL intractable}"] & \textbf{min } \mathrm{MI}_W(\hat{y}, \hat{s}) \arrow[dr, bend left, "\hat{y}\text{ not differentiable}"] & \\
\textbf{min } \mathrm{MI}(\hat{y}, s) \arrow[ur, bend left, "\text{Unobserved sensitive attributes}"] & & & \textbf{min } \mathrm{MI}_W(z_y, z_s)
\end{tikzcd}
$$

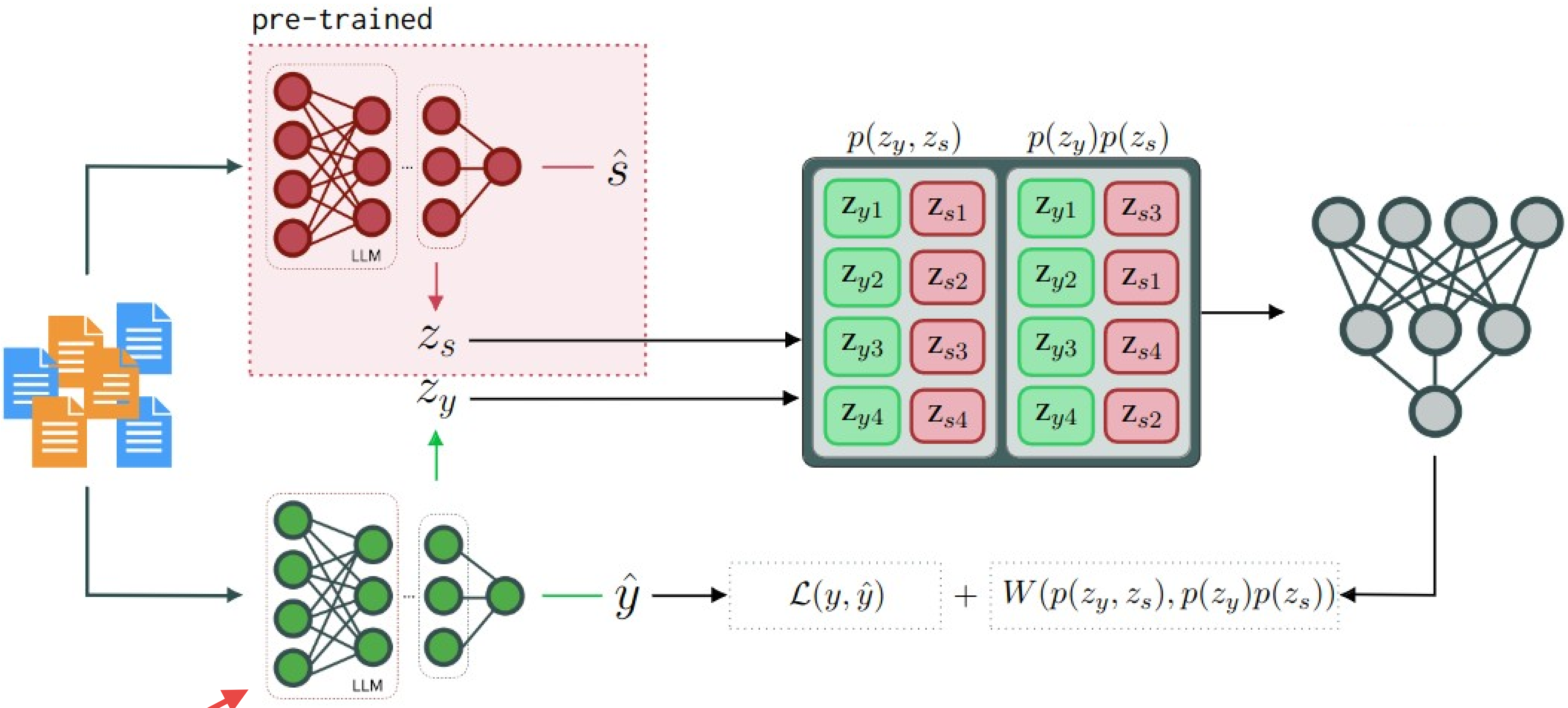
Our approximations upper-bound the original measures:

- $s \rightarrow \hat{s}$, creates an approximation due to the error of the model predicting $\hat{s}$.
- $(\hat{y}, \hat{s}) \rightarrow (z_y, z_s)$ creates a approximation due to the error introduced by the softmax.
- Wasserstein Dependency Measure is related with common fairness metrics.

Architecture

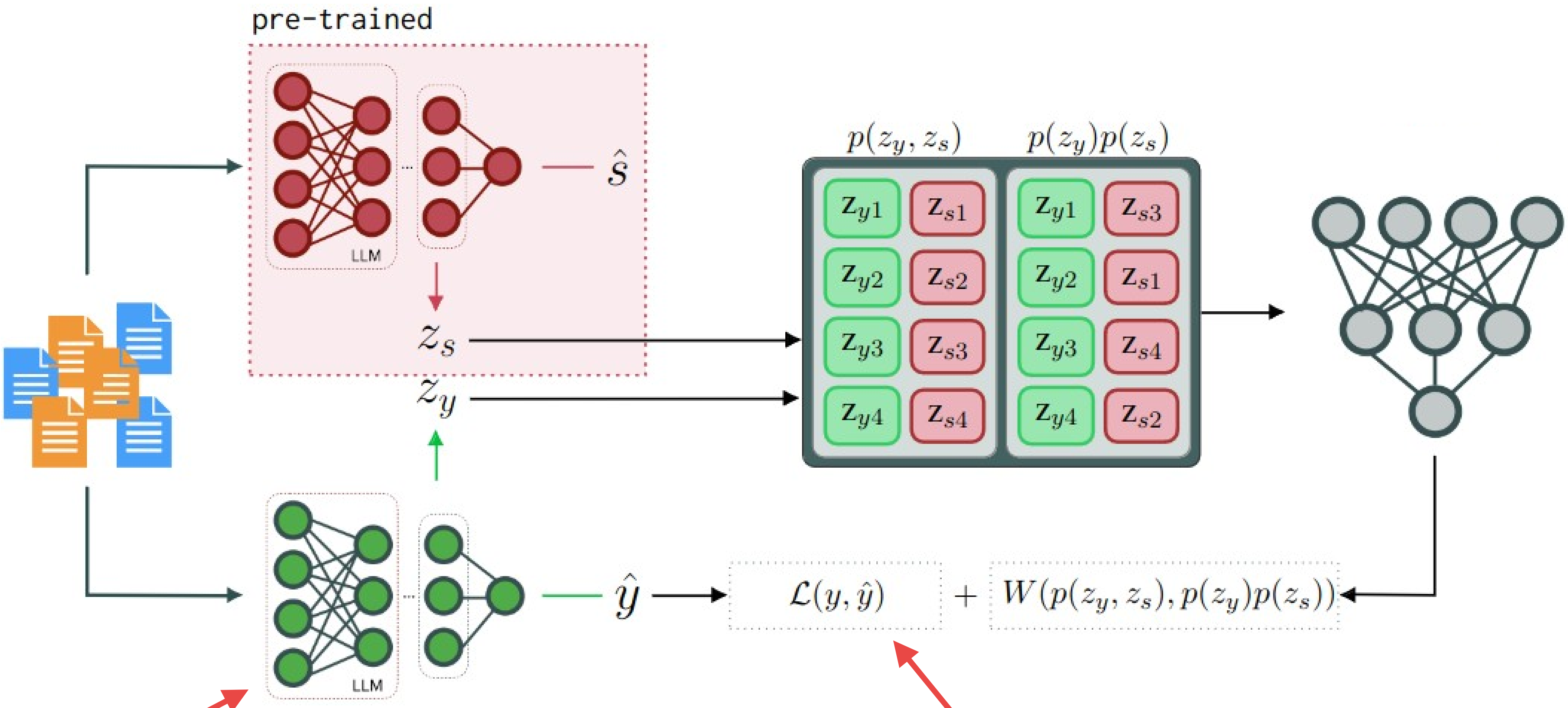
Objective: Training a classification model while minimizing the Wasserstein Dependency Measure between the representations.

Architecture visualization



main model

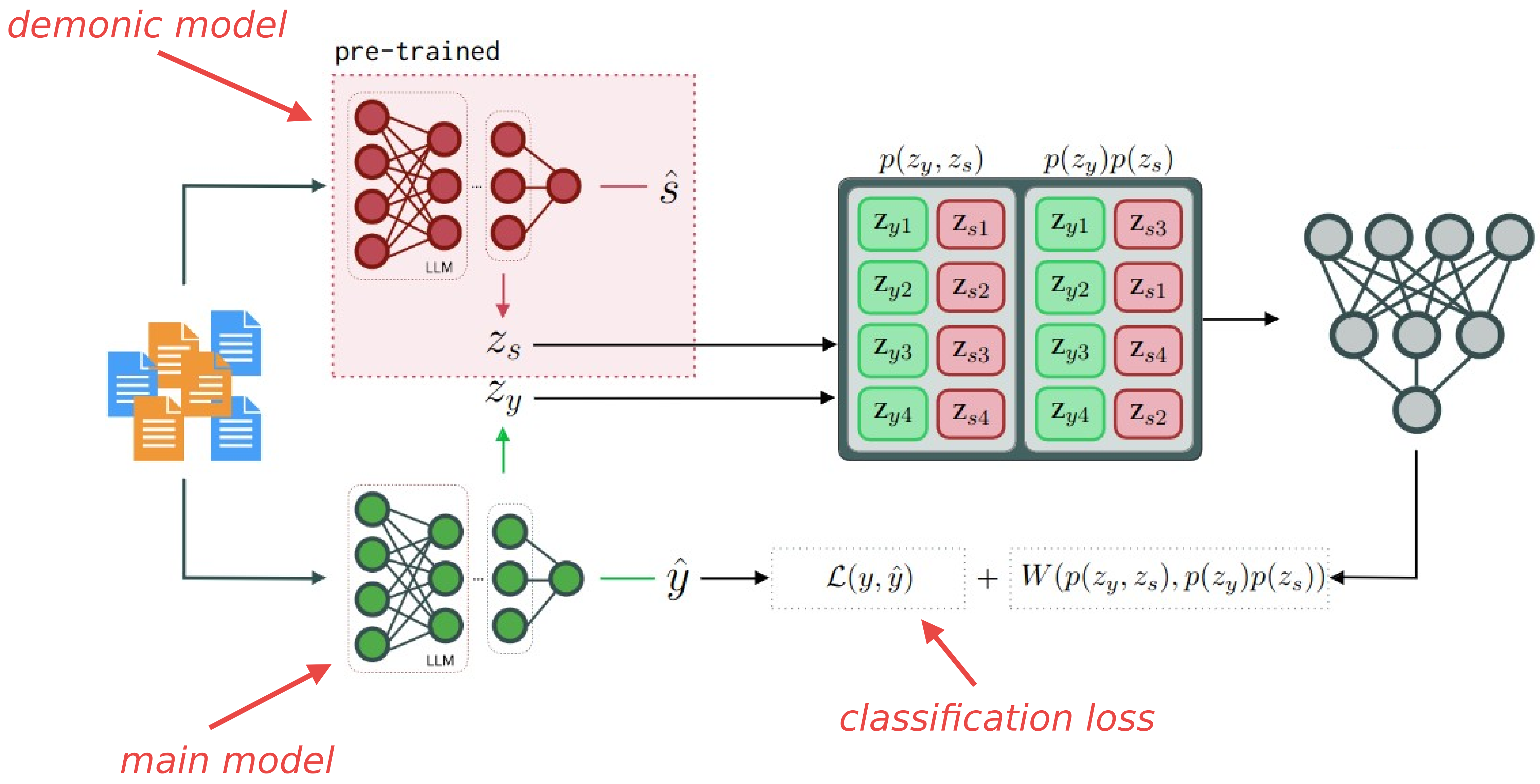
Architecture visualization



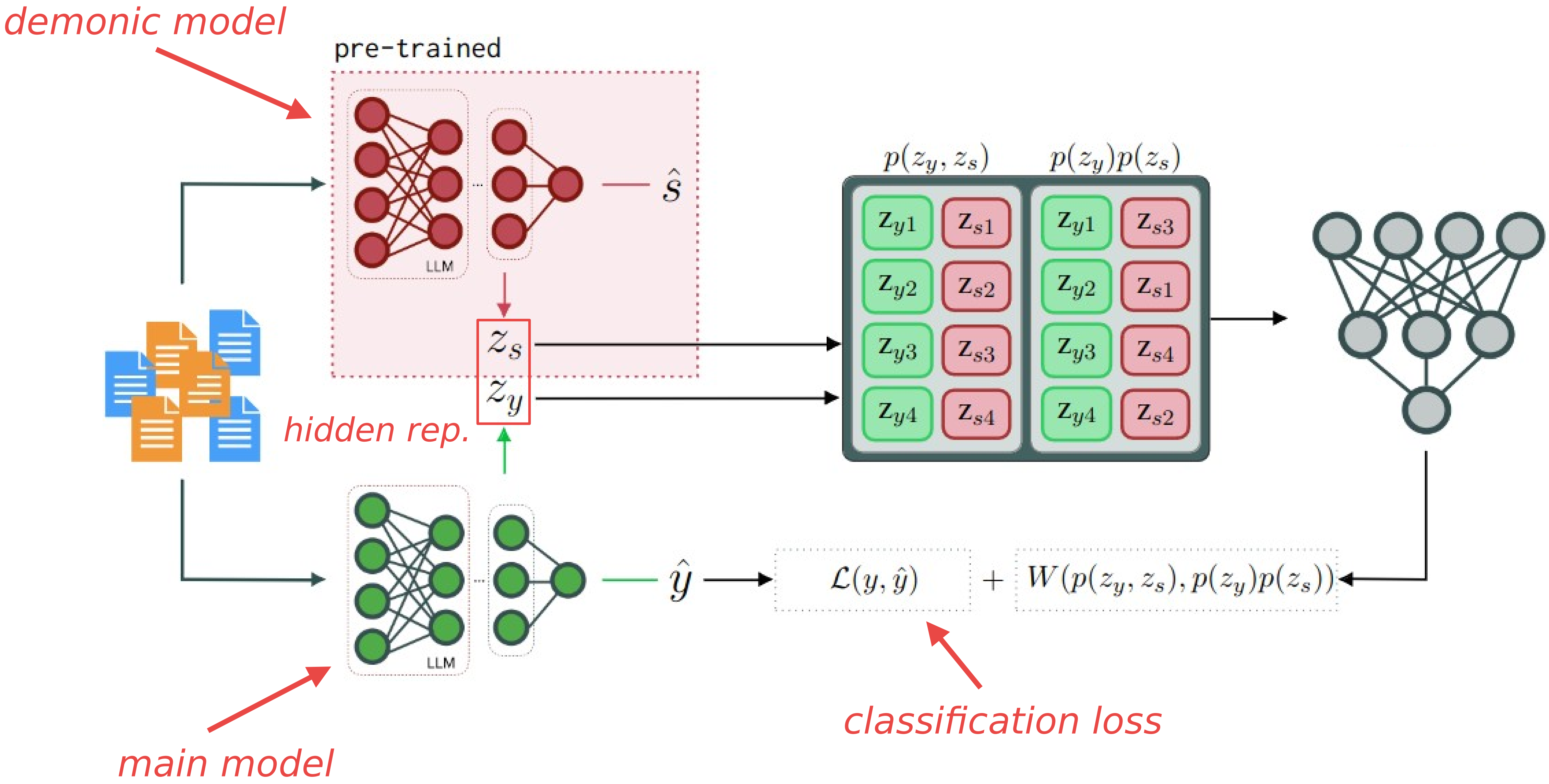
classification loss

main model

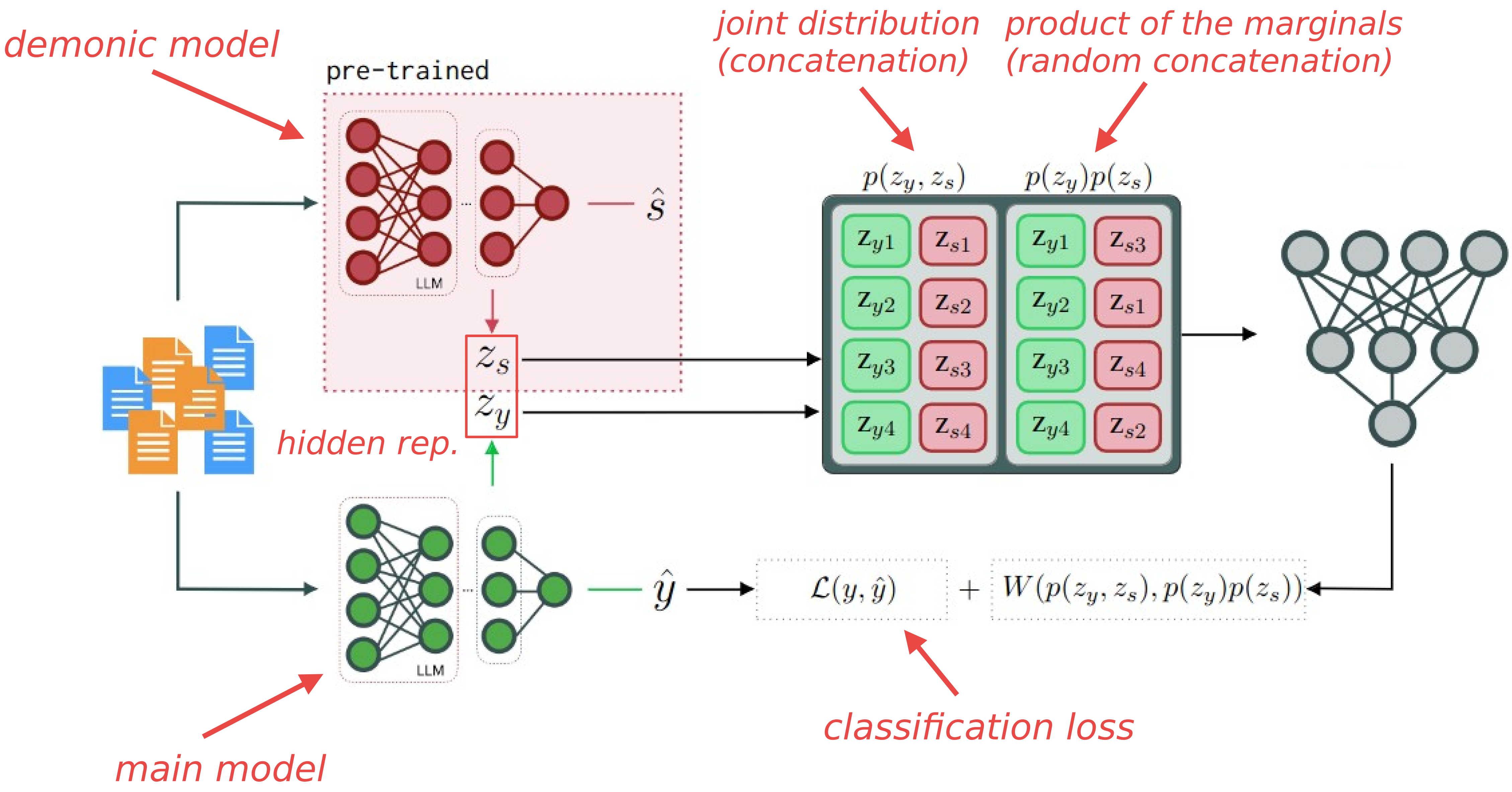
Architecture visualization



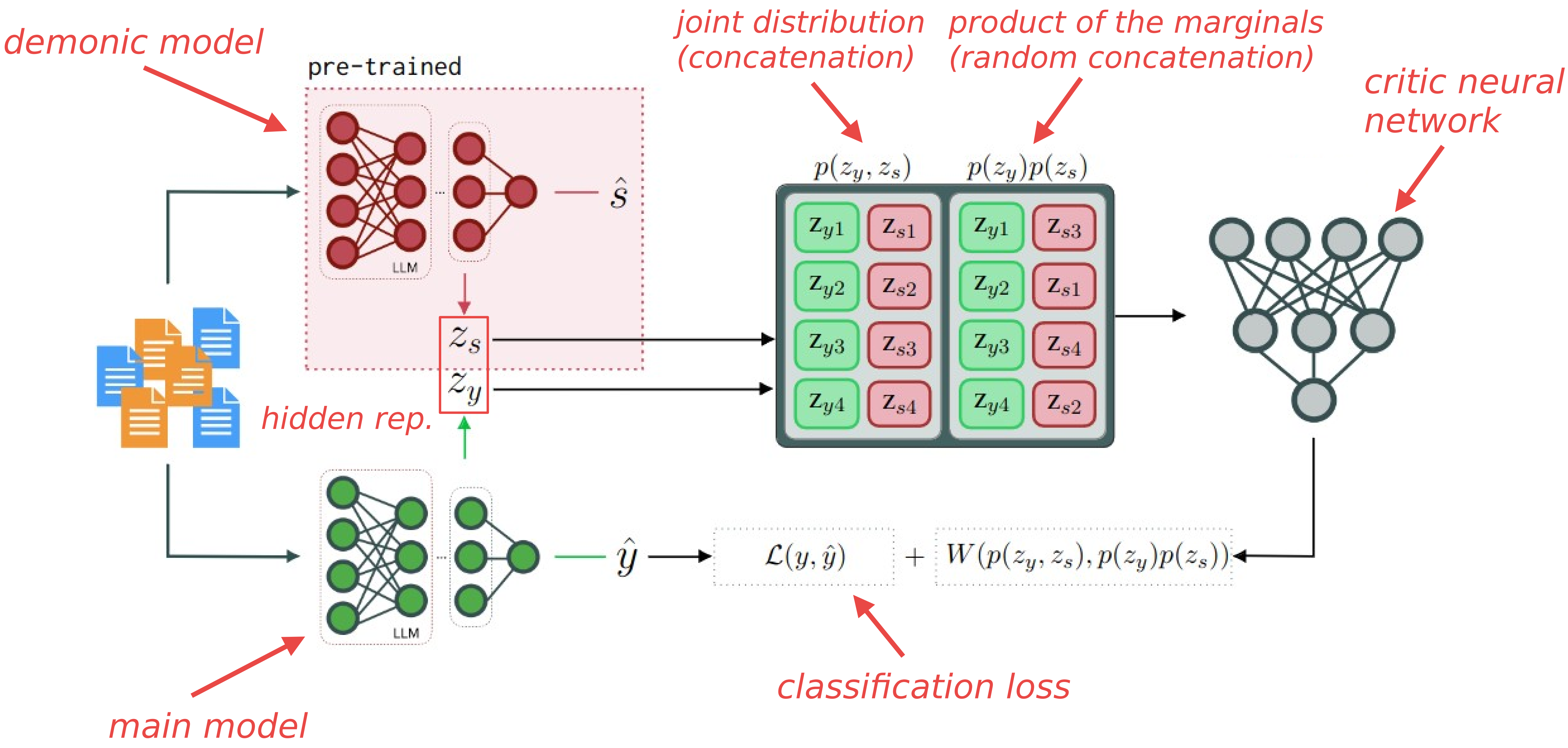
Architecture visualization



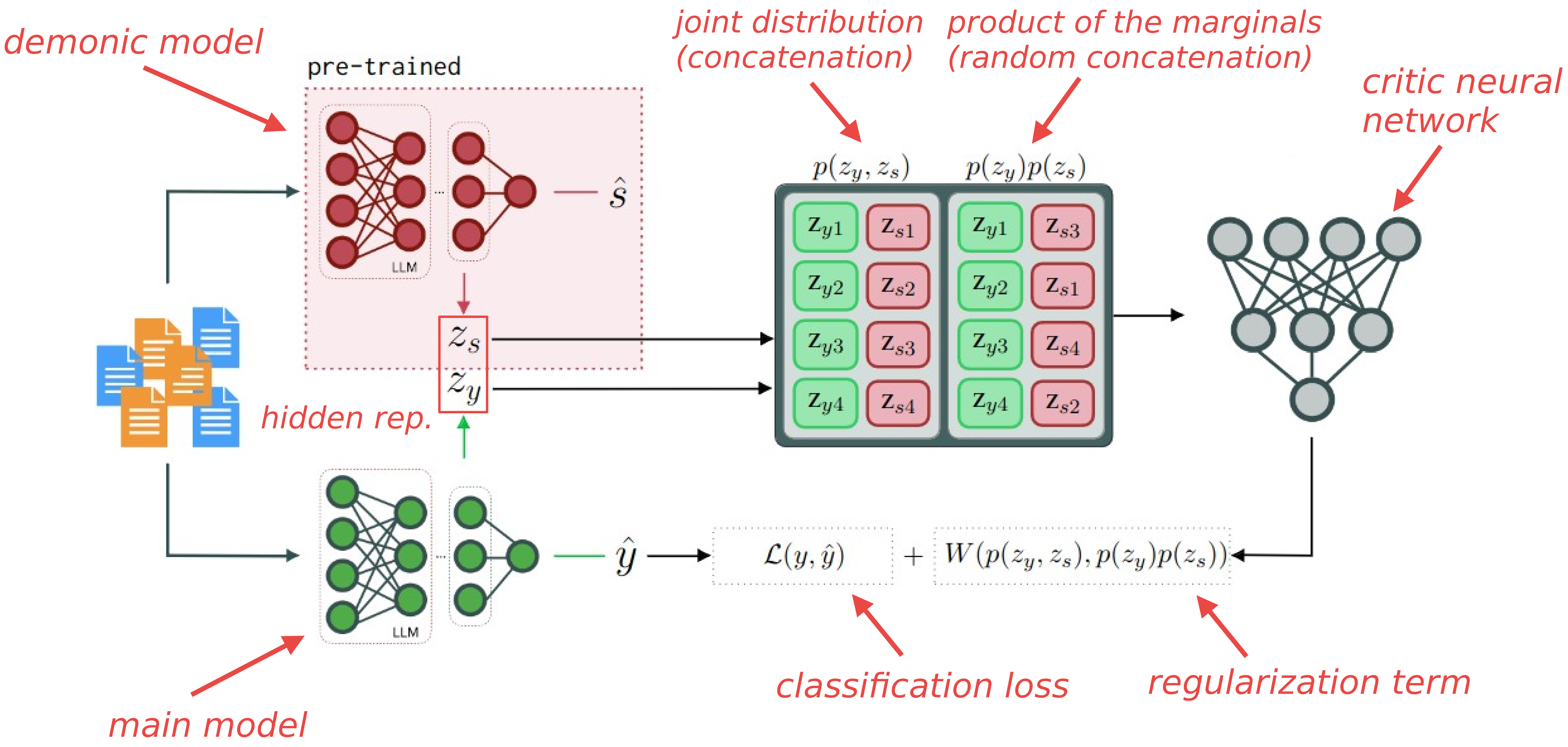
Architecture visualization



Architecture visualization



Architecture visualization



Experiments

Data

1. Bias in Bios dataset (De-Arteaga et al., 2019)

"Mr. Miserez devotes a substantial portion of his practice to representing healthcare entities in the defense of RAC, Medicare, Medicaid and other third party payor audits. He has significant experience defending hospitals, home health agencies, and other healthcare providers (...)"

Label Attorney (occupation)
Sensitive attribute Male (gender)

2. Moji dataset (Blodgett et al., 2016)

"Why am I struggling at work right now"

Label Negative sentiment
Sensitive attribute Standard-American English (type of language)

"Sleep Real Good Cuss Ain Got No Worries"

Label Positive sentiment
Sensitive attribute African-American English (type of language)

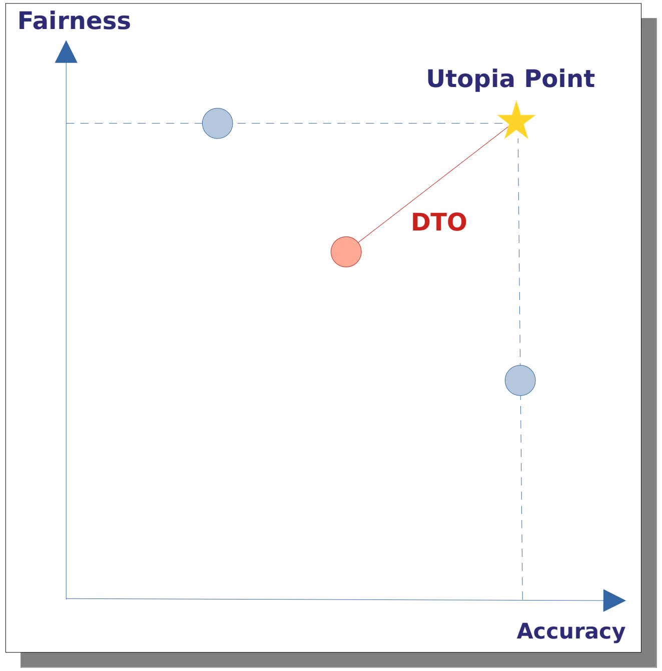
Evaluation framework

Evaluation criteria

Performance on the classification task: accuracy.

Fairness: difference of true positive rate between each sensitive group.

Performance-fairness trade-off: distance to optimum (utopia point).



Main results

Model	Accuracy ↑	Fairness ↑	DTO ↓
*CE	72.3 ± 0.5	61.2 ± 1.4	31.0
INLP	73.3 ± 0.0	85.6 ± 0.0	7.02
Adv	75.6 ± 0.4	90.4 ± 1.1	1.71
Gate	**76.2 ± 0.3**	90.1 ± 1.5	1.90
FairBatch	75.1 ± 0.6	90.6 ± 0.5	1.78
EOGLB	75.2 ± 0.2	90.1 ± 0.4	2.15
Condp	75.8 ± 0.3	88.1 ± 0.6	3.92
Coneo	74.1 ± 0.7	84.1 ± 3.0	8.17
WFC	75.2 ± 0.1	**91.4 ± 0.3**	**1.17**

a) Moji dataset

Model	Accuracy ↑	Fairness ↑	DTO ↓
*CE	82.3 ± 0.2	85.1 ± 0.8	5.67
INLP	82.3 ± 0.0	88.6 ± 0.0	2.44
Adv	81.9 ± 0.2	**90.6 ± 0.5**	1.80
Gate	**83.7 ± 0.2**	90.4 ± 0.9	**0.20**
FairBatch	82.2 ± 0.1	89.5 ± 1.3	1.86
EOGLB	81.7 ± 0.4	88.4 ± 1.0	2.97
Condp	82.1 ± 0.2	84.3 ± 0.8	6.50
Coneo	81.8 ± 0.3	85.2 ± 0.4	5.72
WFC	82.4 ± 0.1	89.0 ± 0.3	2.06

b) Bias in Bios dataset

Table : Model's evaluation. For baselines, results are drawn from (Shen et al., 2022a). We report the mean ± standard deviation over 5 runs. * indicates the model without fairness consideration.

Conclusion

Proposition

Disentanglement of representations to train a fair classifier.

Performance

Competitive or better than state-of-the-art baselines

Advantages

Applicable when datasets lack of sensitive attribute annotations.

Generalizable to other encoder-decoder architecture and to other sensitive attributes (continuous).

Thank you

More details in the paper *Fair Text Classification with Wasserstein Independence*.

Contact

thibaud.leteno@univ-st-etienne.fr
antoine.gourru@univ-st-etienne.fr

DIKÉ project website

Bibliography

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, pages 1630–1640.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In FAccT, pages 610–623.

Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron Van den Oord, Sergey Levine, and Pierre Sermanet. 2019. Wasserstein dependency measure for representation learning. Advances in Neural Information Processing Systems, 32.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In ICML, pages 214–223. PMLR

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. ArXiv preprint arXiv:1608.08868.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In FaccT, pages 120–128.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. ArXiv preprint arXiv:2305.18189.

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems.

Bibliography

Soares, Ioana Baldini, et al. "Your fairness may vary: pretrained language model fairness in toxic text classification." Annual Meeting of the Association for Computational Linguistics. 2022.

Huang, Xiaolei, et al. "Multilingual Twitter Corpus and Baselines for Evaluating Demographic Bias in Hate Speech Recognition." Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020.

Davani, Aida Mostafazadeh, et al. "Improving Counterfactual Generation for Fair Hate Speech Detection." Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021). 2021.

Veale, M., and R. Binns. "Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data." Big Data and Society 4.2 (2017).

Transfer of sensitive attributes

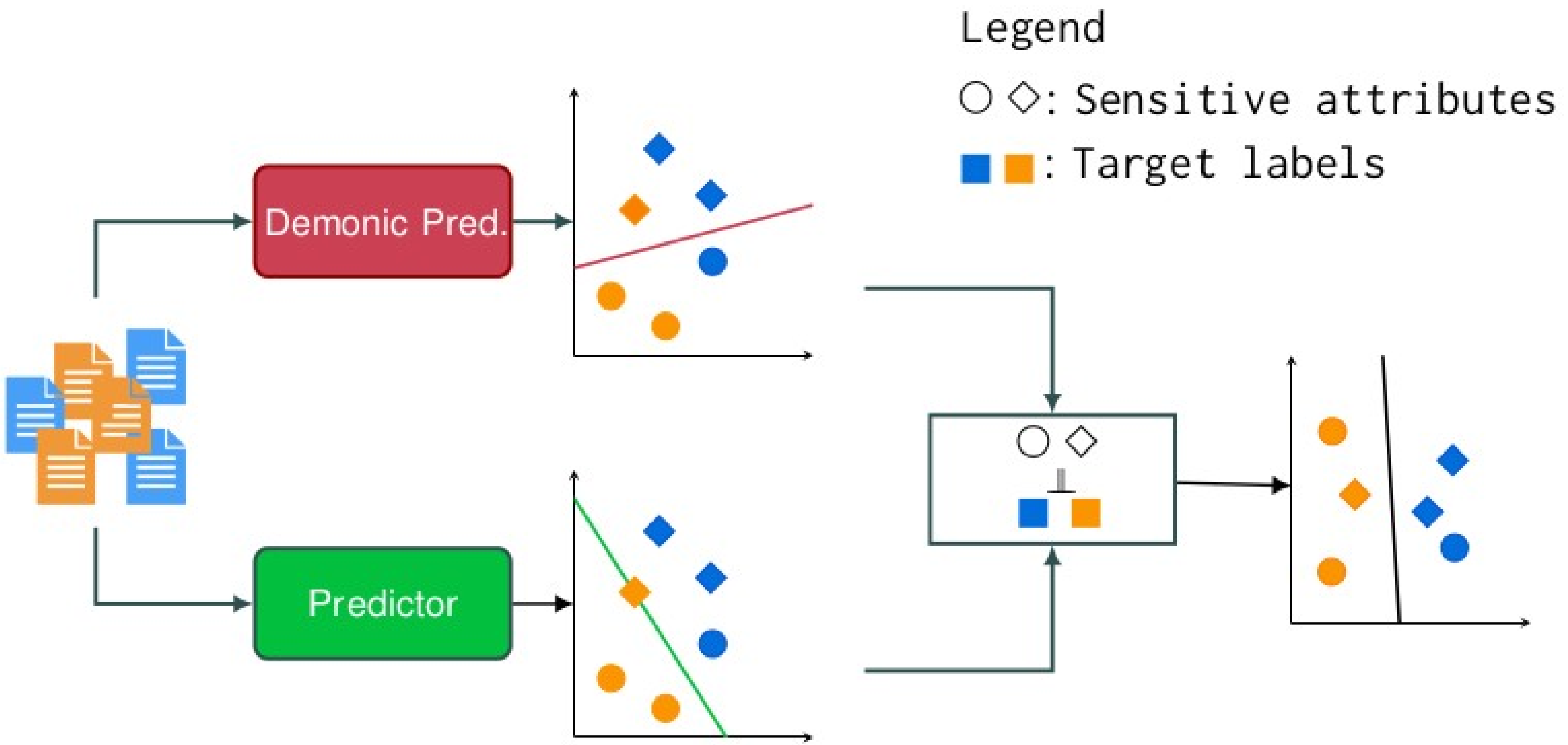


Figure : simplified representation of the method architecture.

Transfer of sensitive attributes

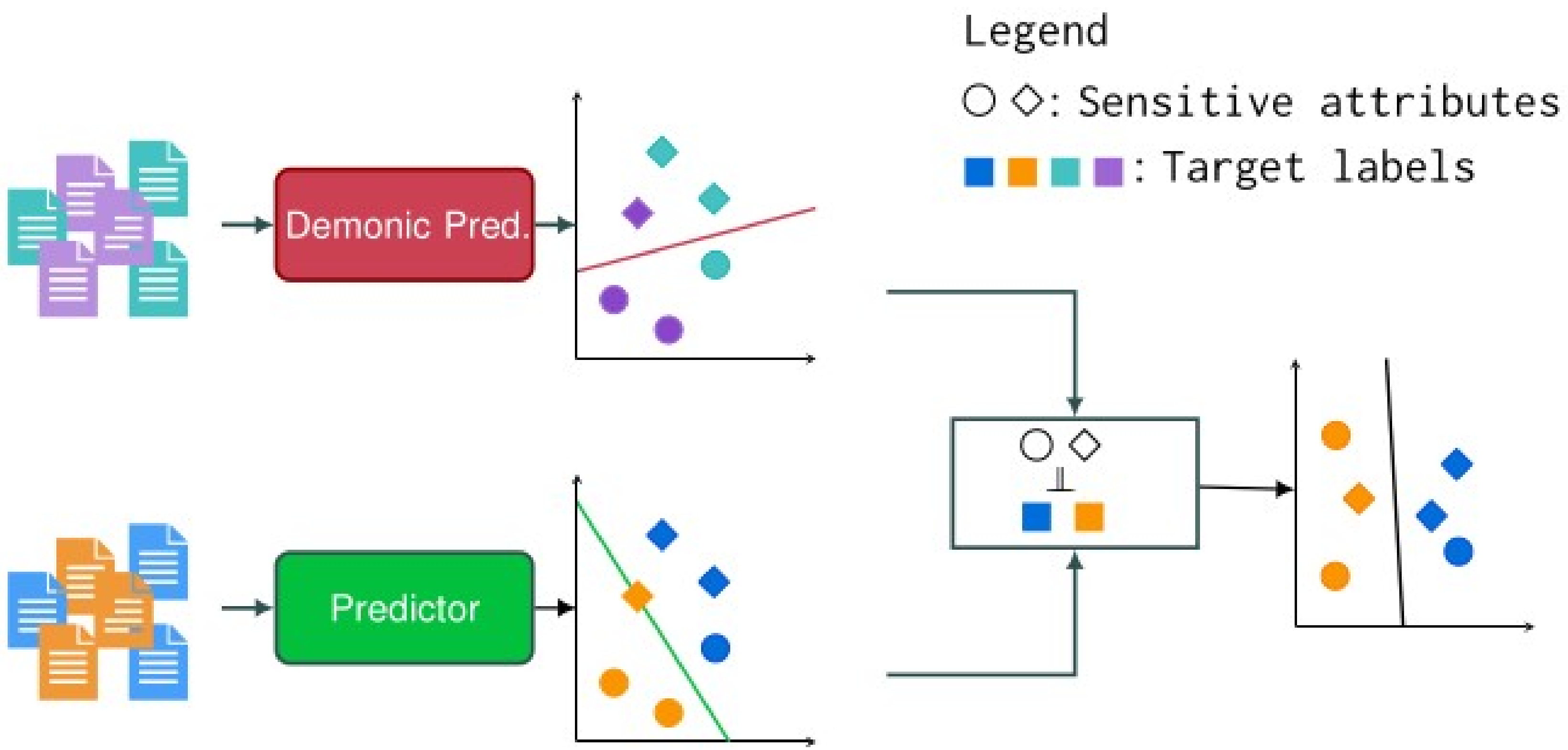


Figure : representation of the transfer of sensitive attributes.

Transfer of sensitive attributes

Dataset	Accuracy ↑	Fairness ↑	DTO ↓	Leakage↓
Bios	**82.4 ± 0.1**	**89.0 ± 0.3**	**2.06**	96.5 ± 0.5
EEC	82.2 ± 0.4	88.9 ± 0.4	2.26	97.5 ± 0.3
MP	**82.4 ± 0.3**	**88.9 ± 0.4**	2.14	**96.4 ± 0.5**

Table : Comparison between several scenarios for training the demonic model for prediction on Bias in Bios.

Other datasets

Equity Evaluation Corpus (EEC): synthetic dataset for sentiment analysis with gender bias. (Kiritchenko and Mohammad, 2018)

Marked Personas dataset (MP): description of personas generated by NLP systems. (Cheng et al., 2023)

Wasserstein dependency measure

min MI($\hat{y}$, $\hat{s}$) → min MI$_W$($\hat{y}$, $\hat{s}$)

Wasserstein distance approximation (Arjovsky et al., 2017)

Use of a neural network called Critic C_ω.

- $W_1(p(\alpha, \beta), p(\alpha)p(\beta)) = \sup_{\omega,\ ||C_\omega||_L \leq 1} \mathrm{E}_{\alpha,\beta \sim p(\alpha,\beta)}[C_\omega(\alpha, \beta)] - \mathrm{E}_{\alpha \sim p(\alpha), \beta \sim p(\beta)}[C_\omega(\alpha, \beta)]$

  where $||C_\omega||_L$ is the set of all 1-Lipschitz functions.

- *Optimization :* $\max \mathrm{E}_{\hat{y},\hat{s} \sim p(\hat{y},\hat{s})}[C_\omega(\hat{y}, \hat{s})] - \mathrm{E}_{\hat{y} \sim p(\hat{y}), \hat{s} \sim p(\hat{s})}[C_\omega(\hat{y}, \hat{s})]$