DIKé project:
some attempts to improve fairness in (compressed) language models

Julien Velcin

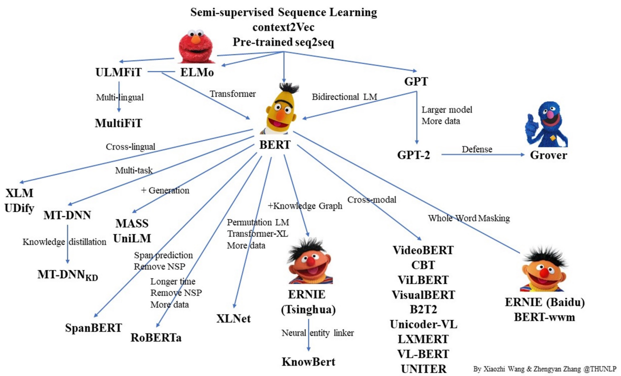
Laboratoire ERIC – Université Lumière Lyon 2

https://eric.univ-lyon2.fr/jvelcin/

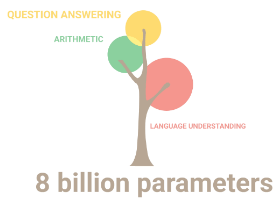
Outline

- Introduction to LLMs, compression and fairness
- Presentation of the DIKé project
- Focus on hate speech detection (work of I. Proskurina)
- Conclusion and ongoing research on LLMs

From pre-trained Language Models…



…to Large Language Models (LLMs)

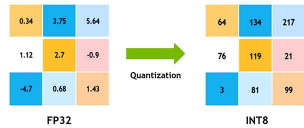


Model compression (1)

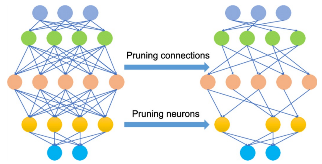
Model	# Parameters
BERT	110M
BERT-large	340M
GPT-2	1.5B
LLaMA	65B
GALACTICA	120B
GPT-3	175B
PaLM	540B

- Transformers are **over-parametrized** (#parameters >> #data)
- Can we just **train smaller** Transformers? No (lottery ticket hypothesis)
- **Mobile** devices

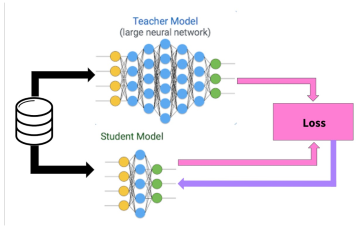
Model compression (2)



iii) Quantization



i) Model pruning



ii) Knowledge distillation

Bias and Fairness

- **Bias**: discrepancy between the correct way of reasoning, which ensures the validity of the conclusions we draw, and the actual process of reasoning. Often, biases arise from the application of heuristics. Some biases can be related to identity features, such as gender or religion.

- **Fairness** in AI: how to correct algorithmic bias in automated decision processes based on machine learning models, *in particular when the biases target groups of people* (e.g., christians).

DIKé project https://www.anr-dike.fr

- Funded by ANR (AAPG 2021, 2022-2025)
- Partners
  - Laboratoire Hubert Curien (LabHC), Université Jean Monnet
  - Laboratoire ERIC, Université Lumière Lyon 2
  - Naver Labs
- Expectations
  - evaluation framework and methodology for evaluating fairness of NLP systems
  - English but also French datasets for fairness and ethics of NLP systems
  - new compressed, fairer language models

Partenaires (PRCE)

LABORATOIRE HUBERT CURIEN
UMR • CNRS • 5516 • SAINT-ETIENNE

Christophe GRAVIER (PR)
François JACQUENET (PR)
Antoine GOURRU (MCF)
Thibaud LETENO (PHD)

Charlotte LACLAU (MCF)
(Télécom Paris)

eric

Julien VELCIN (PR)
Guillaume METLZER (MCF)
Adrien GUILLE (MCF)
Irina Proskurina (PHD)

NAVER LABS Europe

Vassilina NIKOULINA (R. Sc)
Caroline BRUN (Sr Sc.)

Some recent contribution on LLMs

Naver Labs
- SMaLL-100: Introducing Shallow Multilingual Machine Translation Model for Low-Resource Languages (EMNLP 2023)
- What Do Compressed Multilingual Machine Translation Models Forget? (Findings of EMNLP 2023)

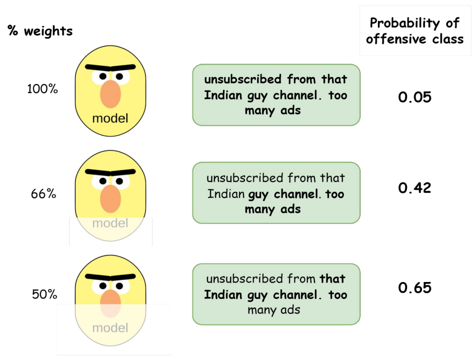
LabHC
- An Investigation of Structures Responsible for Gender Bias in BERT and DistilBERT (IDA 2023)
- Fair Text Classification with Wasserstein Independence (EMNLP 2024)

ERIC
- **The Other Side of Compression: Measuring Bias in Pruned Transformers** (IDA 2023)
- Mini Minds: Exploring Bebeshka and Zlata Baby Models (CoNLL 2023)
- When Quantization Affects Confidence of Large Language Models? (NAACL Findings 2024)

The Other Side of Compression: Measuring Bias in Pruned Transformers (IDA 2023)

- Work of **Irina Proskurina** (PhD student with G. Metzler and me)
- We measure identity-based **bias** in pruned Transformer LMs
- We study **which group** of encoder **layers** (bottom, middle or upper) can be efficiently pruned without biased outcomes
- We propose **word-level supervision** in pruned Transformer LMs as a debiasing method

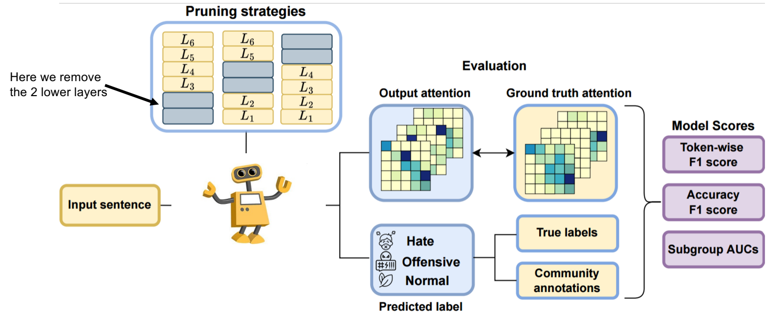
Bias in Hate Speech Classification



Bias = Compressed LM classifies neutral text as offensive and pays ‘attention’ to sensitive attributes

Methodology

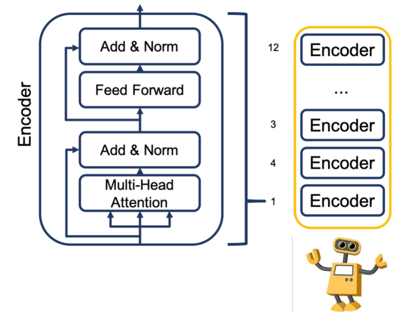
1) Prune Transformer (e.g., BERT)
2) Fine-tune Transformer on hate speech classification task
3) Evaluate performance, bias, and explainability of fine-tuned pruned Transformers



Step 1) Transformers With Pruned Layers (1)

Pruning **K** Layers From Transformer:

- Upper = {12,11,10...}
- Bottom = {1,2,3,...}
- Symmetric = {6,7}
- Alternate Odd = {11,9, 7, ...}
- Alternate Even = {12,10,8...}
- Contribution*-based



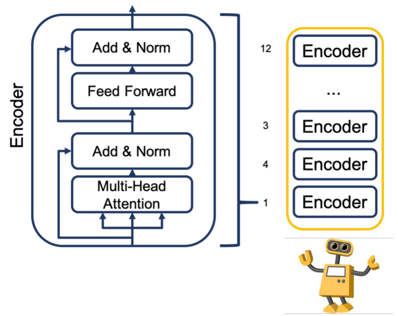
*Measured with cosine similarity of hidden states: $\varphi_s(l) = \cos(Z_{l-1}, Z_l)$

Step 1) Transformers With Pruned Layers (2)

Pruning **K** Layers From Transformer:

- LMs: BERT, RoBERTa, DistiBERT, DistilRoBERTa
- **K** ={2,4,6} for BERT-based LMs
- **K** ={1,2,3} for Distilled LMs
- Contribution*-based strategy:

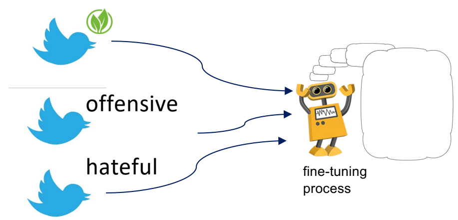
  BERT: {5, 10, 9, 7, 2, 4}
  RoBERTa: {1, 2, 6, 8, 9, 4}
  DistilBERT: {2, 3, 4}
  DistilRoBERTa: {6, 2, 3}



*Measured with cosine similarity of hidden states: $\varphi_s(l) = \cos(Z_{l-1}, Z_l)$

Step 2) Fine-tune Transformer on hate speech classification task

- Given input text, Transformer is fine-tuned to classify text as offensive, hateful or neutral
- Data: HATEXPLAIN
- Fine-tuning: Cross-Entropy loss $L(\theta) = -\sum_{i=1}^{c} y_i \log(\hat{y}_i)$, where $i \in$ {offensive, hateful, neutral}, y_i is true class, $\hat{y}_i$ output model probability of i belongs to one of the classes



Step 3) Evaluate Bias in Compressed models (1)

- Data divided in 4 domains: $D_t^+, D_t^-, D_{\setminus t}^+, D_{\setminus t}^-$
- t – (hate) target community, +/- : class, neutral or hateful
- Subgroup AUC = AUC(D_t^+, D_t^-)
- Background Positive Subgroup Negative: BPSN = AUC($D_{\setminus t}^+$, D_t^-)
- Background Negative Subgroup Positive: BNSP = AUC($D_{\setminus t}^-$, D_t^+)

Subgroup AUC

documents related to the group t

d	pred	GT
12	0.76	hateful
2	0.76	neutral
307	0.73	neutral
17	0.72	hateful
1	0.69	neutral
101	0.61	neutral
...		

Step 3) Evaluate Explainability in Compressed models (2)

- Explainability performance: token-level F1-score
- F1-score(output attention, ground truth attention)
- Output attentions=top-5 tokens with higher attention weights
- Ground truth attention = annotations in red

<user>: I got a guilty pleasure and it is country music and hillbilly movies and tv shows about rednecks hunting in the woods... trailer[ab] trash[abc] poor[c] plump[c] thing[c]

[a]Annotator 1: Target labels: *Economic, Caucasian*
[b]Annotator 2: Target labels: *Economic*
[c]Annotator 3: Target labels: *Caucasian*

Rationales
= words targeting a given group

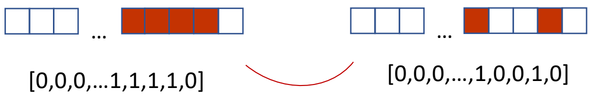
Step 3) Evaluate Explainability in Compressed models (3)

<user>: I got a guilty pleasure and it is country music and hillbilly movies and tv shows about rednecks hunting in the woods... trailer[ab] trash[abc] poor[c] plump[c] thing[c]

[a]Annotator 1: Target labels: *Economic, Caucasian*
[b]Annotator 2: Target labels: *Economic*
[c]Annotator 3: Target labels: *Caucasian*

True Rationales

Predicted Rationales = Attention Weights > 0.2



[0,0,0,...1,1,1,1,0]

[0,0,0,...,1,0,0,1,0]

Token-level F1-score

Step 3) Evaluate Bias in Compressed models (4)

If the impact of compression is uniform, then the shift in scores achieved on the texts mentioning a target community *t* should also be uniform compared to the overall scores shift. That forms our null hypothesis H0.

$$H_0 : \beta_0^t - \beta_0 = \beta_c^t - \beta_c$$

no significant difference

$$H_1 : \beta_0^t - \beta_0 \neq \beta_c^t - \beta_c,$$

significant difference

Results: Compressed LMs are prone to bias

full model

4 layers removed

Model	Layers	F1 score	Token F1 score	Count Signif Target Classes		
				Subgroup	BNSP	BPSN
BERT	12/12	$67.28_{\pm 0.13}$	$48.58_{3.28}$	-	-	-
	10/12	$65.31_{\pm 0.17}$	$38.35_{\pm 4.11}$	2	0	1
	8/12	$64.82_{\pm 0.15}$	$32.57_{\pm 4.06}$	2	0	2
	6/12	$63.46_{\pm 0.21}$	$34.4_{\pm 3.87}$	4	0	2
DistilBERT	6/6	$66.19_{\pm 0.44}$	$43.31_{\pm 3.42}$	-	-	-
	5/6	$66.08_{\pm 0.62}$	$42.77_{\pm 4.13}$	0	0	0
	4/6	$65.66_{\pm 0.51}$	$42.1_{\pm 3.98}$	3	0	1
	3/6	$64.31_{\pm 0.83}$	$39.81_{\pm 4.22}$	3	1	2
RoBERTa	12/12	$83.42_{\pm 0.4}$	$46.64_{\pm 3.51}$	-	-	-
	10/12	$81.46_{\pm 0.41}$	$39.37_{\pm 4.61}$	4	2	2
	8/12	$78.67_{\pm 0.58}$	$38.49_{\pm 4.23}$	6	3	4
	6/12	$77.08_{\pm 0.33}$	$24.47_{\pm 4.08}$	6	5	5
DistilRoBERTa	6/6	$82.02_{\pm 0.36}$	$42.08_{\pm 5.24}$	-	-	-
	5/6	$81.08_{\pm 0.4}$	$33.2_{\pm 4.75}$	3	0	2
	4/6	$77.06_{\pm 0.48}$	$32.76_{\pm 5.21}$	3	2	4
	3/6	$74.05_{\pm 0.43}$	$32.6_{\pm 4.61}$	6	5	6

$\cancel{H_0}$ H_1

number of groups with a significant difference in term of classification (on 10)

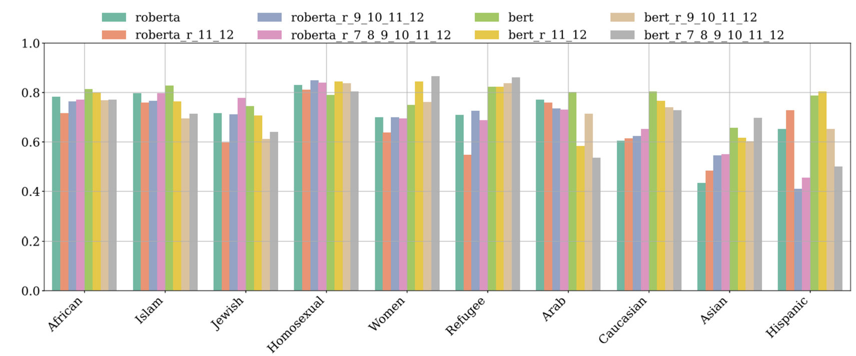
Performance of original and pruned models on HATEXPLAIN test set

Results: Compressed LMs rely on unimportant tokens

Model	Layers	F1 score	Token F1 score	Count Signif Target Classes		
				Subgroup	BNSP	BPSN
BERT	12/12	$67.28_{\pm 0.13}$	$48.58_{3.28}$	-	-	-
	10/12	$65.31_{\pm 0.17}$	$38.35_{\pm 4.11}$	2	0	1
	8/12	$64.82_{\pm 0.15}$	$32.57_{\pm 4.06}$	2	0	2
	6/12	$63.46_{\pm 0.21}$	$34.4_{\pm 3.87}$	4	0	2
DistilBERT	6/6	$66.19_{\pm 0.44}$	$43.31_{\pm 3.42}$	-	-	-
	5/6	$66.08_{\pm 0.62}$	$42.77_{\pm 4.13}$	0	0	0
	4/6	$65.66_{\pm 0.51}$	$42.1_{\pm 3.98}$	3	0	1
	3/6	$64.31_{\pm 0.83}$	$39.81_{\pm 4.22}$	3	1	2
RoBERTa	12/12	$83.42_{\pm 0.4}$	$46.64_{\pm 3.51}$	-	-	-
	10/12	$81.46_{\pm 0.41}$	$39.37_{\pm 4.61}$	4	2	2
	8/12	$78.67_{\pm 0.58}$	$38.49_{\pm 4.23}$	6	3	4
	6/12	$77.08_{\pm 0.33}$	$24.47_{\pm 4.08}$	6	5	5
DistilRoBERTa	6/6	$82.02_{\pm 0.36}$	$42.08_{\pm 5.24}$	-	-	-
	5/6	$81.08_{\pm 0.4}$	$33.2_{\pm 4.75}$	3	0	2
	4/6	$77.06_{\pm 0.48}$	$32.76_{\pm 5.21}$	3	2	4
	3/6	$74.05_{\pm 0.43}$	$32.6_{\pm 4.61}$	6	5	6

Performance of original and pruned models on HATEXPLAIN test set

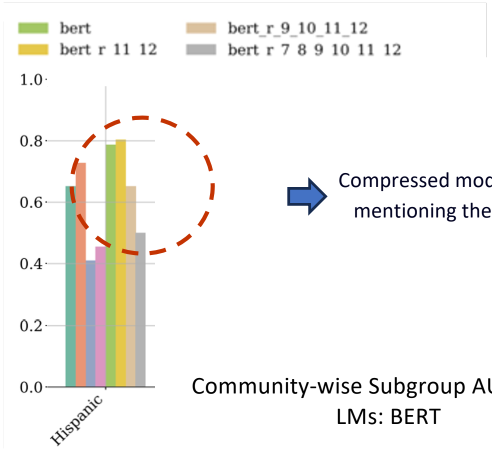
Results: The impact of compression is not uniform



OY: Subgroup AUC scores on HateXplain, OX: Target communities

LMs: BERT, RoBERTa

Results: The impact of compression is not uniform



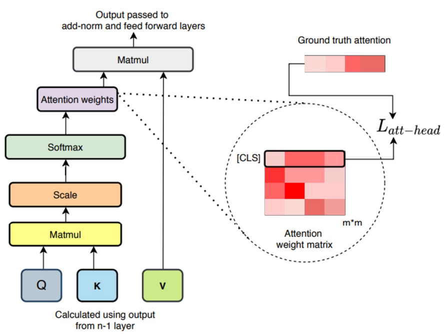
Compressed models classify neutral texts mentioning the community as offensive/hateful

Community-wise Subgroup AUC scores

LMs: BERT

Solution: Supervised Attention learning

$$Loss_{\sum} = Loss_{pred} + \lambda Loss_{attn}$$



<user>: I got a guilty pleasure and it is country music and hillbilly movies and tv shows about rednecks hunting in the woods... trailer[ab] trash[abc] poor[c] plump[c] thing[c]

[a]Annotator 1: Target labels: *Economic, Caucasian*
[b]Annotator 2: Target labels: *Economic*
[c]Annotator 3: Target labels: *Caucasian*

True Rationales

[0,0,0,...1,1,1,1,0]

Predicted Rationales

[0,0,0,...0.25,0,0,0.3,0..16,0]

Results: Fine-tuning with attention loss compensates for fairness loss

Model	λ	F1 score	Token F1 score	Subgroup AUC
BERT (6/12)	0	$63.46_{\pm 0.21}$	$34.4_{\pm 3.87}$	$0.59_{\pm 0.01}$
	0.01	$65.12_{\pm 0.38}$	$36.3_{\pm 4.01}$	$0.707_{\pm 0.11}$
	0.1	$65.92_{\pm 0.24}$	$39.26_{\pm 3.91}$	$0.784_{\pm 0.07}$
	1	$66.61_{\pm 0.17}$	$45.54_{\pm 3.29}$	$0.803_{\pm 0.12}$
DistilBERT (3/6)	0	$64.31_{\pm 0.83}$	$39.81_{\pm 4.22}$	$0.768_{\pm 0.24}$
	0.01	$64.35_{\pm 0.51}$	$40.4_{\pm 3.04}$	$0.748_{\pm 0.16}$
	0.1	$65.11_{\pm 0.7}$	$41.03_{\pm 3.28}$	$0.794_{\pm 0.31}$
	1	$66.71_{\pm 0.22}$	$42.67_{\pm 3.14}$	$0.796_{\pm 0.28}$
RoBERTa (6/12)	0	$77.08_{\pm 0.33}$	$24.47_{\pm 4.08}$	$0.519_{\pm 0.21}$
	0.01	$80.86_{\pm 0.22}$	$33.19_{\pm 3.28}$	$0.612_{\pm 0.29}$
	0.1	$78.58_{\pm 0.23}$	$36.49_{\pm 4.11}$	$0.681_{\pm 0.17}$
	1	$82.38_{\pm 0.26}$	$40.52_{\pm 3.81}$	$0.691_{\pm 0.14}$
DistilRoBERTa (3/6)	0	$71.05_{\pm 0.43}$	$32.6_{\pm 4.61}$	$0.62_{\pm 0.08}$
	0.01	$79.14_{\pm 0.47}$	$34.41_{\pm 4.11}$	$0.634_{\pm 0.04}$
	0.1	$81.25_{\pm 0.33}$	$36.51_{\pm 3.5}$	$0.635_{\pm 0.08}$
	1	$81.96_{\pm 0.51}$	$43.02_{\pm 4.14}$	$0.65_{\pm 0.09}$

*λ = 0 - non-supervised attention learning

$$Loss_{\sum} = Loss_{pred} + \lambda Loss_{attn}$$

Performance and fairness scores (Subgroup AUC) of models trained with word-level supervision

BERT Subgroup AUC scores
- .59 - without attention supervision
- .80 - with attention supervision

Conclusion on this work

- We conducted two chains of experiments to analyze the effect of Transformer LMs **pruning** in the context of **hate speech classification** tasks (with and without attention supervision)
- We compare **both fairness and performance loss** for pruned BERT, RoBERTa, and their distilled versions
- We show and statistically prove that **removing any layer** from Transformer LMs **results in fairness loss** even when the performance loss could be negligible
- We conduct supervised attention-learning experiments that help to **reduce bias in pruned models**

Recent and ongoing work on LLMs (1)

- Recent work @ERIC:
  - Participation to the **BabyLM challenge** (https://babylm.github.io), shared task of CoNLL 2023. Can we pretrain a model from few data? What kind of task can we still solve?
    Our contribution: “Mini Minds: Exploring Bebeshka and Zlata Baby Models”
  - Work on quantization of LLMs (GPTQ) and **confidence**
    Is there an impact of quantization on the model confidence?
    Our contribution: “When Quantization Affects Confidence of Large Language Models?” (NAACL Findings 2024)

Recent and ongoing work on LLMs (2)

- Ongoing work:
  - French dataset related to **ethical questions**, semi-automatic translation of Moral Stories (Emelin et al., 2021), joint work with LabHC
  - French dataset for **implicit hate speech detection,** joint work with Naver Labs
  - new methods for **debiasing** compressed generative (causal) models
  - study the fairness of LMs in **detecting subtle forms** of hate speech (joint work with C. Favre)

How fair are language models in detecting subtle forms of hate speech? (1)

- Create a set of synthetically generated sentences
  - Chat-GPT generated sentences following instructions based on (manual) patterns such as <They> <are good at> <object> and paraphrased in various ways

  e.g.: *The girls are marketing ninjas. Even under pressure, the girls maintain their composure.*

- Evaluate the difference of toxicity scores for various groups of people
  - Difference between the toxicity score for « They » and the plural noun (e.g. « The girls »)
  - Compute the mean for different models fine-tuned on toxicity (e.g., Hate-BERT)

How fair are language models in detecting subtle forms of hate speech? (2)

- Some results:

Attribute	Sensitivity		Toxic Samples	
	BERT	*RoBERTa*	*BERT*	*RoBERTa*
girls	0.093	0.005	44	2
women	0.070	0.008	44	5
boys	0.098	0.006	71	2
men	0.073	0.006	50	4
buddhists	0.112	0.007	20	2
muslims	0.101	0.013	24	9
jews	0.093	0.052	26	30
christians	0.169	0.064	128	32
atheists	0.143	0.027	110	12

Take-away message

- Transformer-based language models **can be used** for detecting hate speech in texts
- language models **can be biased** when detecting hate speech: neutral sentence with phrase « Indian guy » is classified as hateful by LMs
- **lack of datasets in French** for hate speech detection (we are working on it)
- **Compression can amplify bias** in LMs used for hate speech detection (even though biases are already present in pre-compressed models)
- **Bias can be mitigated** in language models with forcing models not to pay attention to gender/nationality/religion expressions.

Some references

[1] Kenton, Jacob Devlin Ming-Wei Chang, and Lee Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of NAACL-HLT*, pp. 4171-4186. 2019.
[2] Hooker, Sara, et al. "What do compressed deep neural networks forget?." *arXiv preprint arXiv:1911.05248* (2019)
[3] Mathew, Binny, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. "Hatexplain: A benchmark dataset for explainable hate speech detection." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, pp. 14867-14875. 2021.
[4] Gupta, Manish, Vasudeva Varma, Sonam Damani, and Kedhar Nath Narahari. "Compression of deep learning models for NLP." In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 3507-3508. 2020.

Thank you

Email: Julien.Velcin@univ-lyon2.fr
Website: https://eric.univ-lyon2.fr/jvelcin/
DIKé project website: https://www.anr-dike.fr

Projet DIKÉ