Calibration of Natural Language Understanding Models with Venn–ABERS Predictors

Patrizio Giovannotti

ROYAL HOLLOWAY UNIVERSITY OF LONDON

centrica

COPA 2022

NLU vs NLP

Traditional Natural Language Processing

- *N*-grams
- Syntax / Grammars
- POS tagging
- Tokenization
- Information retrieval
- Manually built resources

Natural Language Understanding

- Meaning of a phrase
- Focused on supervised learning

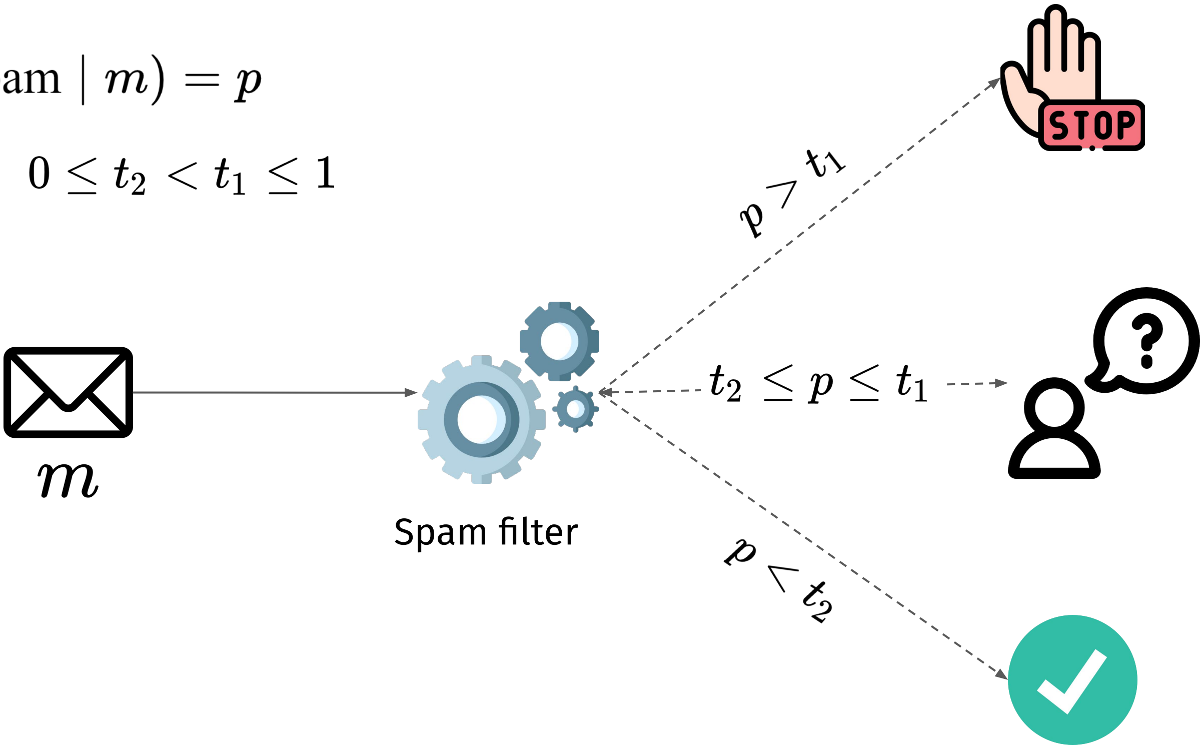
Areas:

- Paraphrase detection
- Language Inference
- Sentiment analysis
- Linguistic acceptability

Uncertainty in NLU: why?

$$P(Y = \text{spam} \mid m) = p$$

Thresholds: $0 \leq t_2 < t_1 \leq 1$



m

Spam filter

$p > t_1$

$t_2 \leq p \leq t_1$

$p < t_2$

Calibration

Input $X \in \mathcal{X}$

Label $Y \in \mathcal{Y} = \{1, \ldots, K\}$

prediction $\hat{Y}$ with prob. $\hat{P}$

$$\mathbb{P}(\hat{Y} = Y \mid \hat{P} = p) = p$$

$\forall p \in [0, 1]$

Example:

Out of all predictions with probability *P*=0.75, about 75% must be correct

Sharpness

- On a balanced test dataset, a classifier always predicting 1 with *P*=0.5 would still be well calibrated
- Term mainly used in forecasting
  - Implies narrow confidence intervals

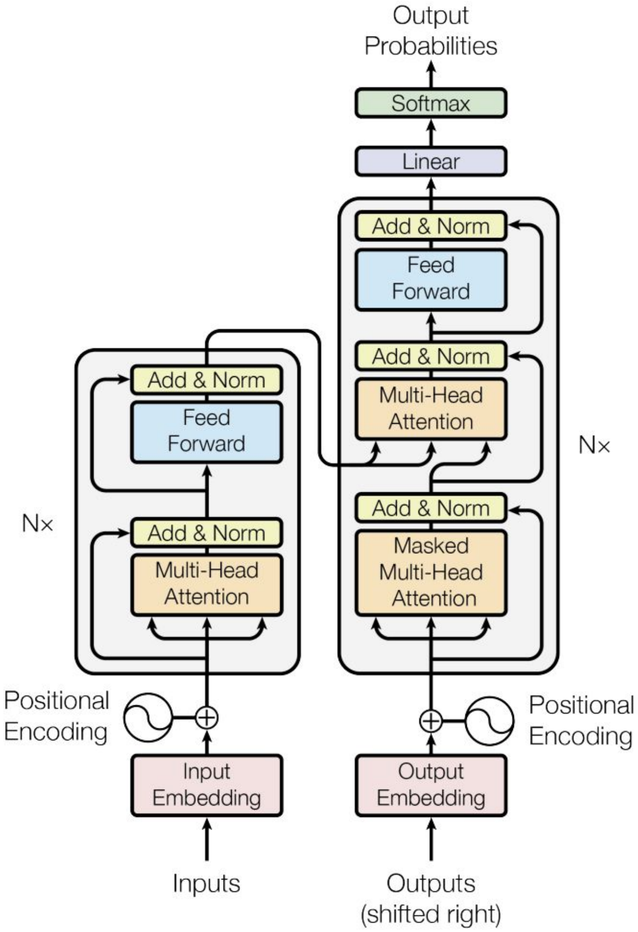
Alternatively:

- A model is sharp if probability estimates are adjusted for each instance
- The distribution of probabilities generated is uniform over [0,1]

State-of-the-art in NLU: Transformers

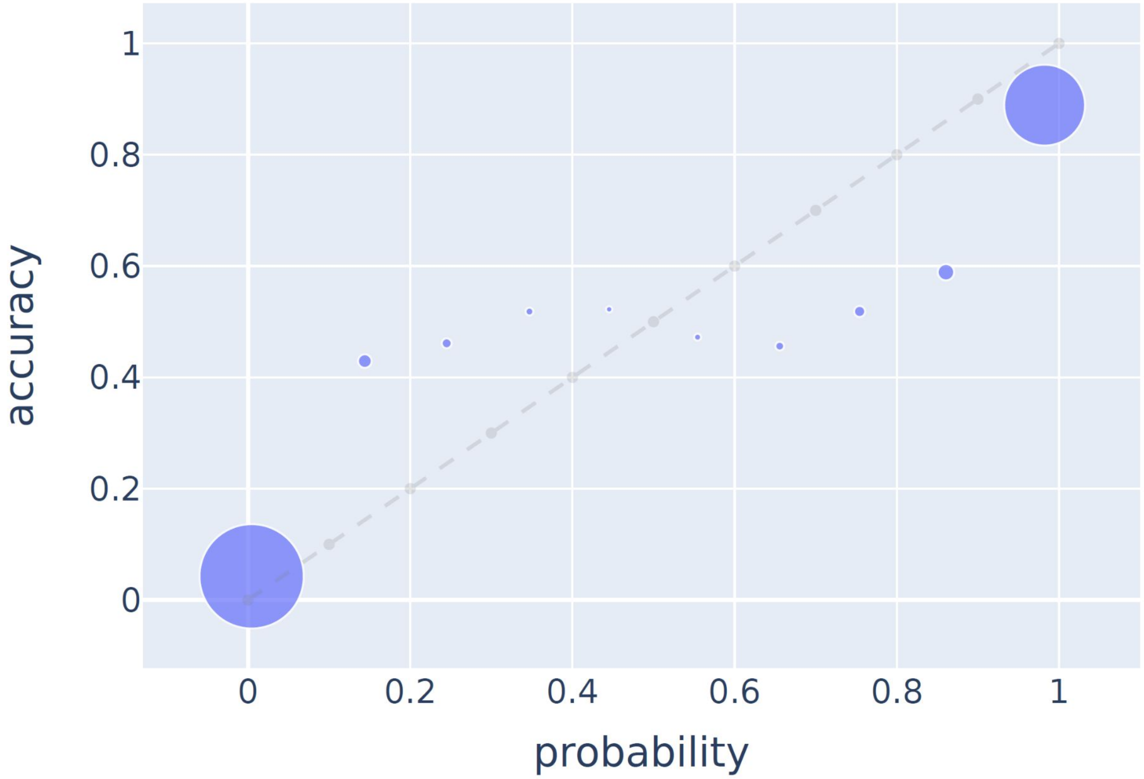
- Born as sequence-to-sequence modellers for machine translation (2017)
- BERT (2018) is based on a transformer encoder
- Many variations, constantly growing in size and performance
- They can come already pre-trained

- They output a “score” for each label
- The score is usually turned into probability via softmax



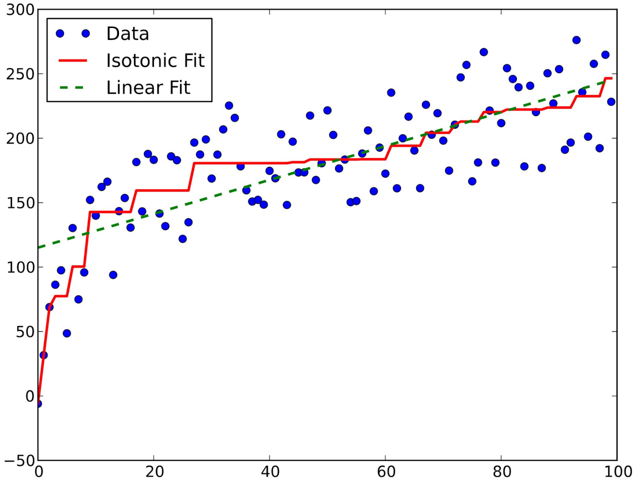
Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

Calibration and sharpness of a RoBERTa model trained on Quora Question Pairs



Venn–ABERS predictors (1)

- Multiprobabilistic predictors
- Built on top of any existing *scoring algorithm*
- Perfectly calibrated (if data is exchangeable)



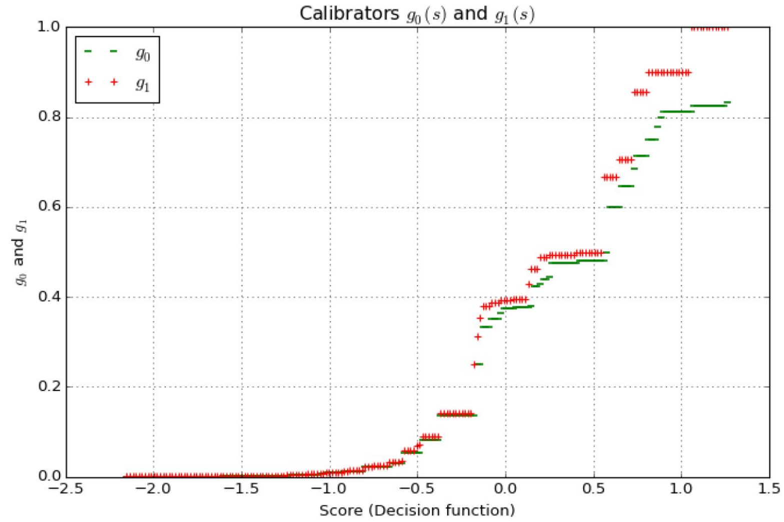
- Inputs *(x, y)*
- Algorithm outputs score *s(x)*
- Find function *g* : *g(s(x))* is a calibrated probability
- *g* is calculated by isotonic regression

Venn–ABERS predictors (2)

- Test input *(x, y)*
- Fit g_0 for *y=0* and g_1 for *y=1*
- Obtain two probabilities, p_0 and p_1

Want only one probability?

$$p = \frac{p_1}{1 - p_0 + p_1}$$



Toccaceli, Paolo, et al. "Excape Wp1. Probabilistic Prediction." (2016).

Datasets (1)

Quora Question Pairs		
How is air traffic controlled?	How do you become an air traffic controller?	0
How should I make myself brave?	How can I be more brave?	1

Stanford Sentiment Treebank	
Steven Spielberg brings us another masterpiece.	0.986
Ultimately feels empty and unsatisfying, like swallowing a Communion wafer without the wine.	0.111

Linguistic Acceptability Corpus	
The building is tall and wide.	0
The building is tall and tall.	1

Datasets (2)

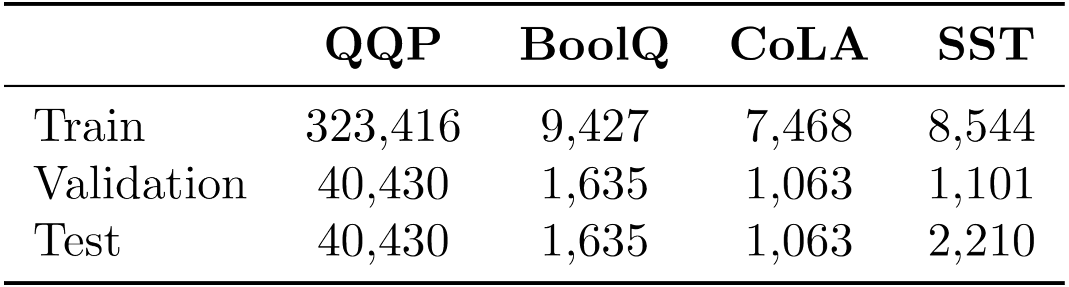
Boolean Questions		
Air Force One -- The Air Force usually does not have fighter aircraft escort the presidential aircraft over the United States but it has occurred, for example during the attack on the World Trade Center.	Does Air Force One travel with fighter escort?	0
Calcium carbide -- Calcium carbide is a chemical compound with the chemical formula of CaC. Its main use industrially is in the production of acetylene and calcium cyanamide.	calcium carbide cac2 is the raw material for the production of acetylene	1

Pretrained models

	# of params (millions)
BERT (2018)	110
RoBERTa (2019)	125
ALBERT (2019)	11
DeBERTa (2021)	143

Training setup

- Dataset splits:



	QQP	BoolQ	CoLA	SST
Train	323,416	9,427	7,468	8,544
Validation	40,430	1,635	1,063	1,101
Test	40,430	1,635	1,063	2,210

- Scores averaged over 5 training trials for all datasets, except QQP
- 3 epochs

Expected Calibration Error (ECE)

- Divide predictions in *M* bins of equal width
- B_m contains examples with confidence ranging in $\left(\frac{m-1}{M}, \frac{m}{M}\right]$

$$\mathrm{ECE} = \frac{1}{n}\sum_{m=1}^{M} |B_m| \cdot |p(B_m) - \hat{p}(B_m)|$$

true fraction of positive instances in bin

average estimated probability for predictions in bin

F1 score (macro)

For a label *k*

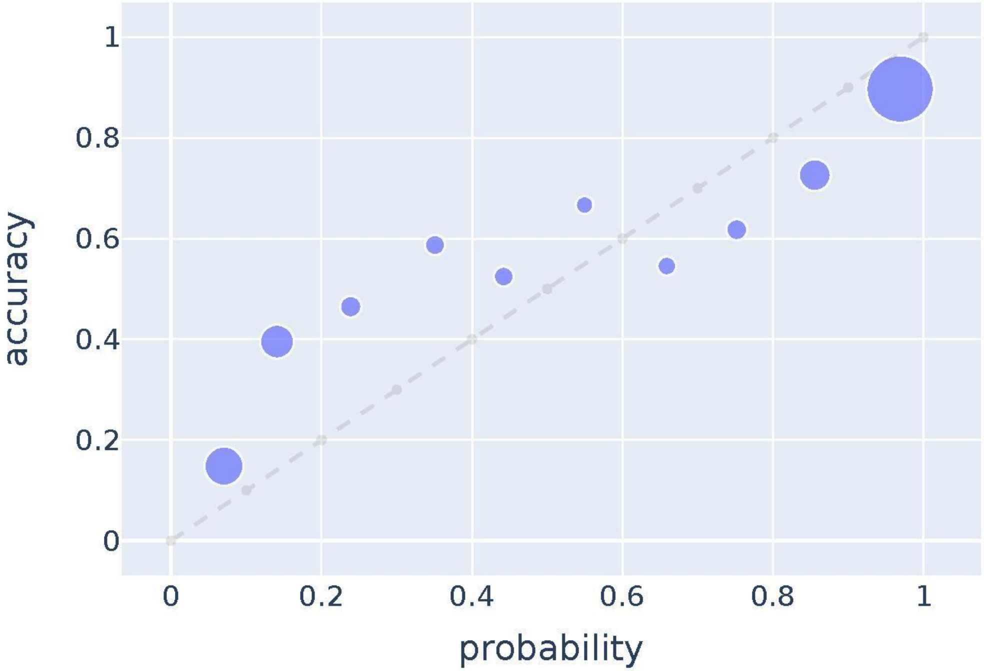
$$F_1^{(k)} = \frac{2P^{(k)}R^{(k)}}{P^{(k)}+R^{(k)}}$$

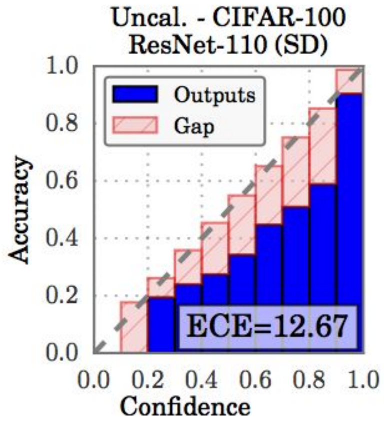
Precision

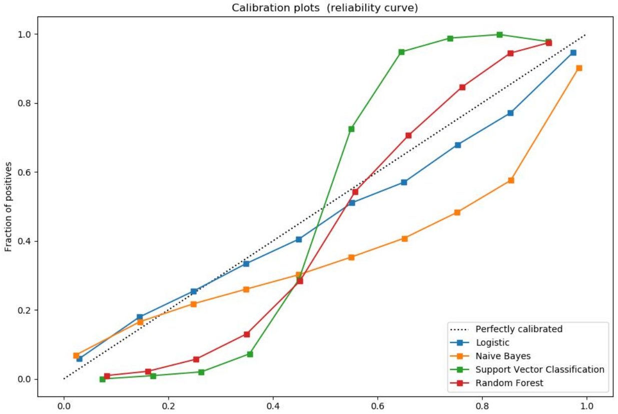
Recall

Macro F1: average over all *k* labels

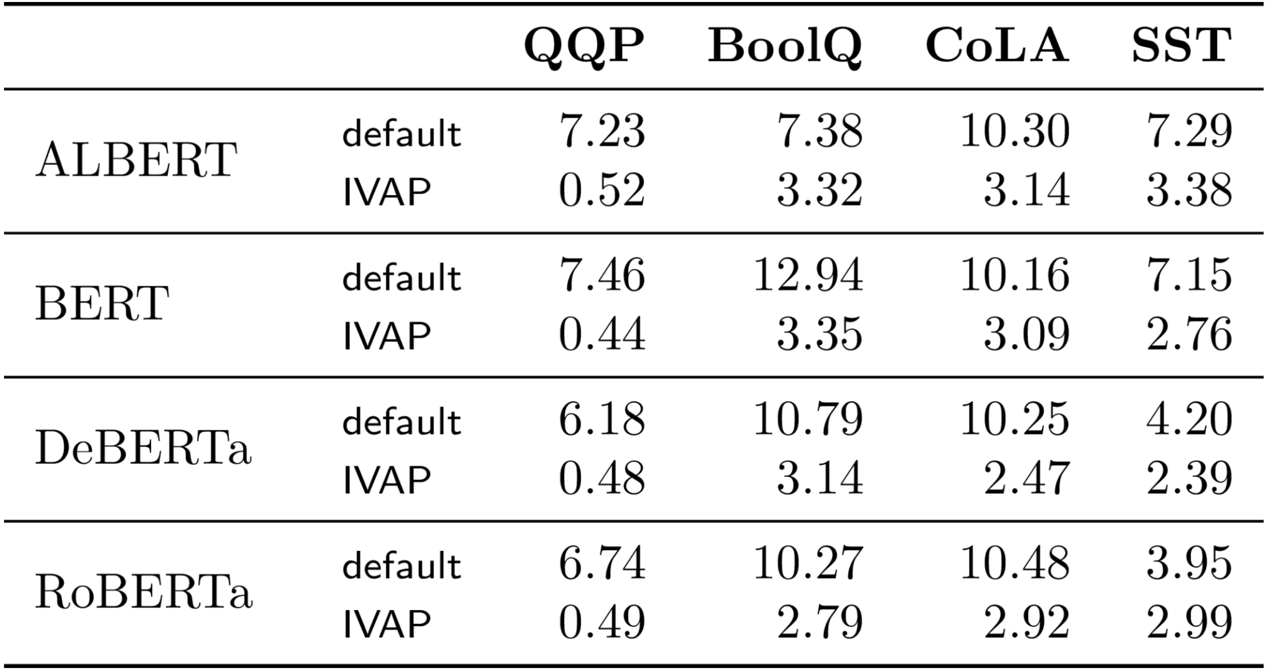
Reliability bubble chart





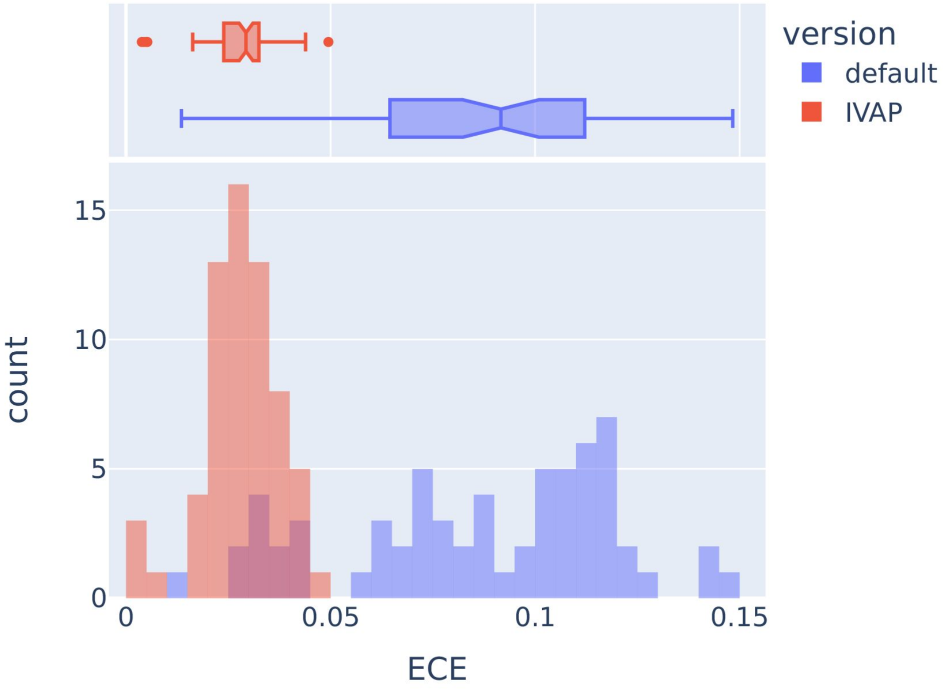


Results



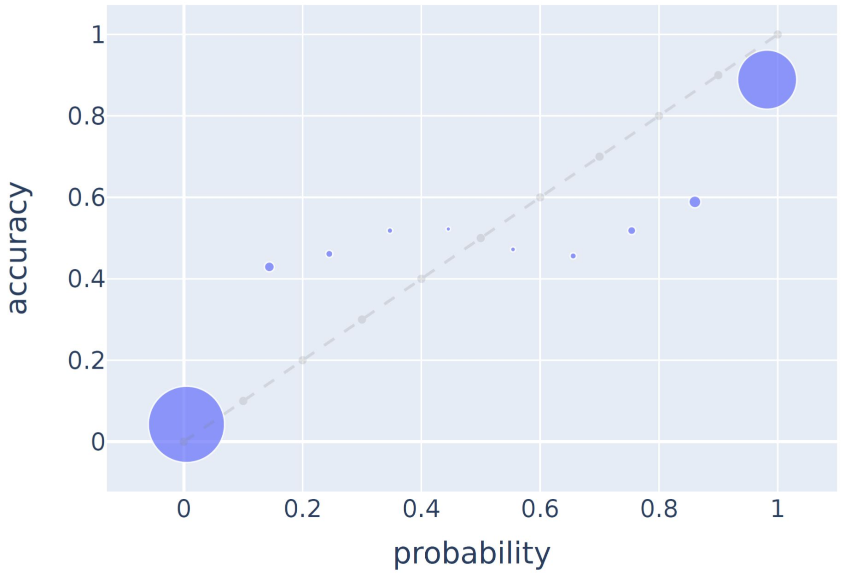
		QQP	**BoolQ**	**CoLA**	**SST**
ALBERT	default	7.23	7.38	10.30	7.29
	IVAP	0.52	3.32	3.14	3.38
BERT	default	7.46	12.94	10.16	7.15
	IVAP	0.44	3.35	3.09	2.76
DeBERTa	default	6.18	10.79	10.25	4.20
	IVAP	0.48	3.14	2.47	2.39
RoBERTa	default	6.74	10.27	10.48	3.95
	IVAP	0.49	2.79	2.92	2.99

Expected calibration error (in %) for default and IVAP models.

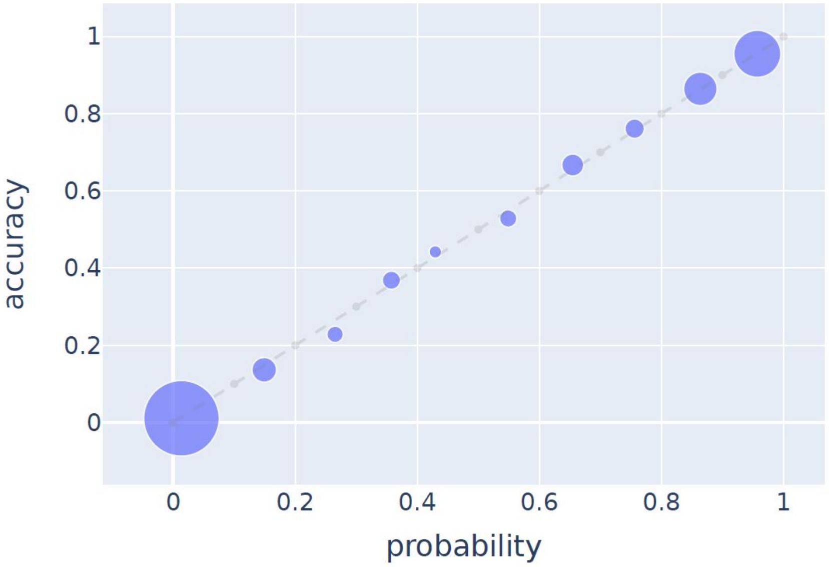


Distribution of expected calibration errors over all datasets, models and trials.

Sharpness: RoBERTa on QQP

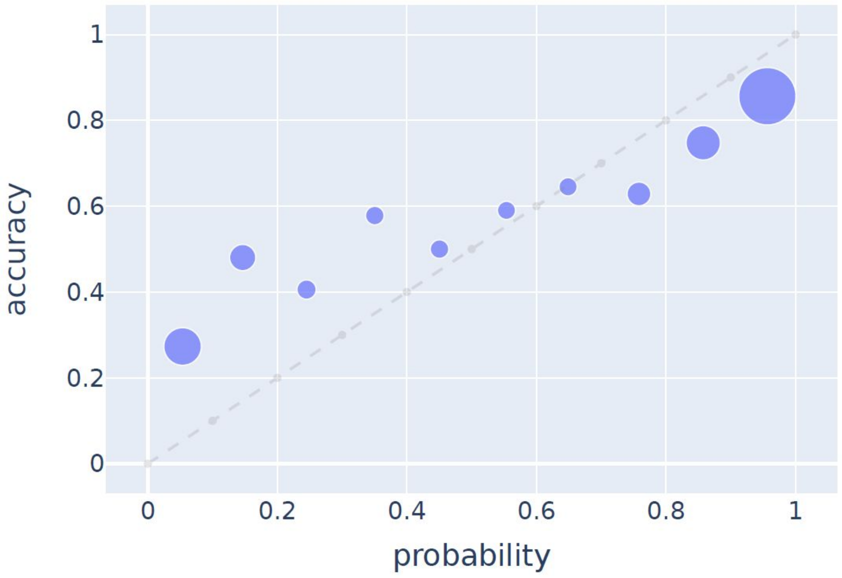


Default

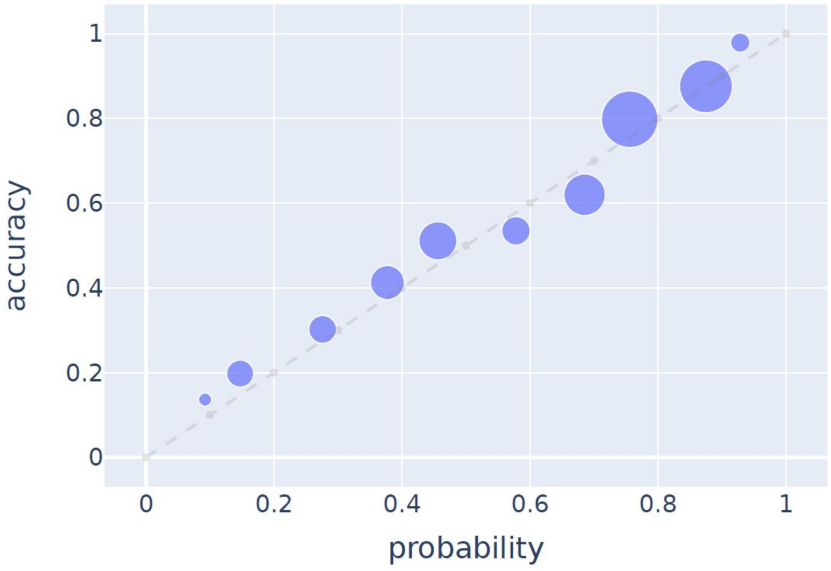


Venn–ABERS

Sharpness: BERT on BoolQ

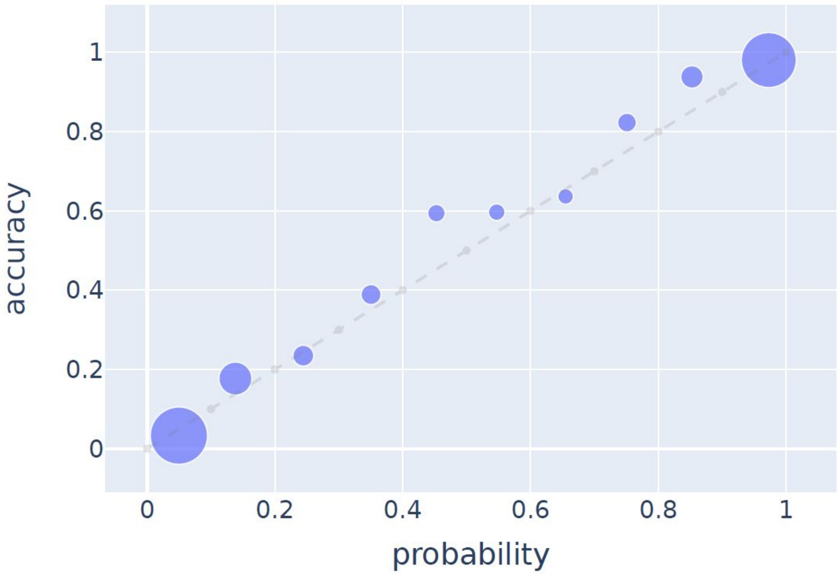


Default

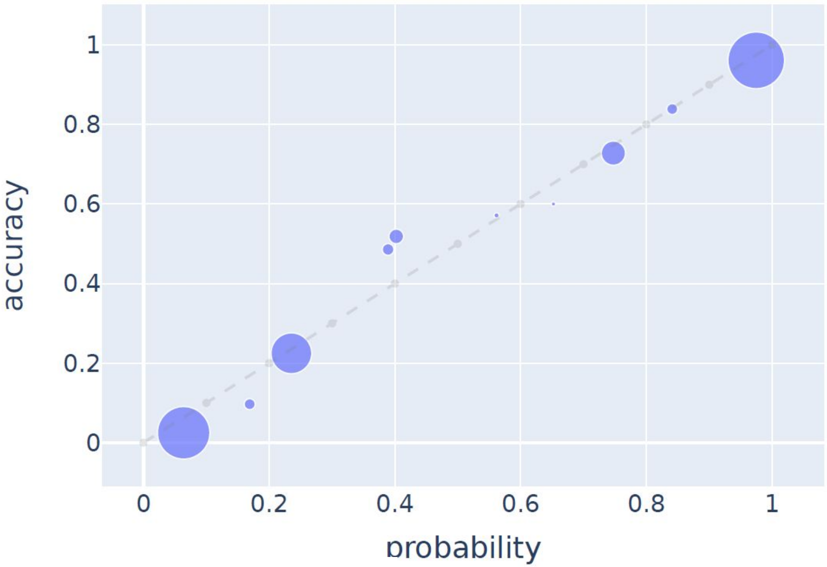


Venn–ABERS

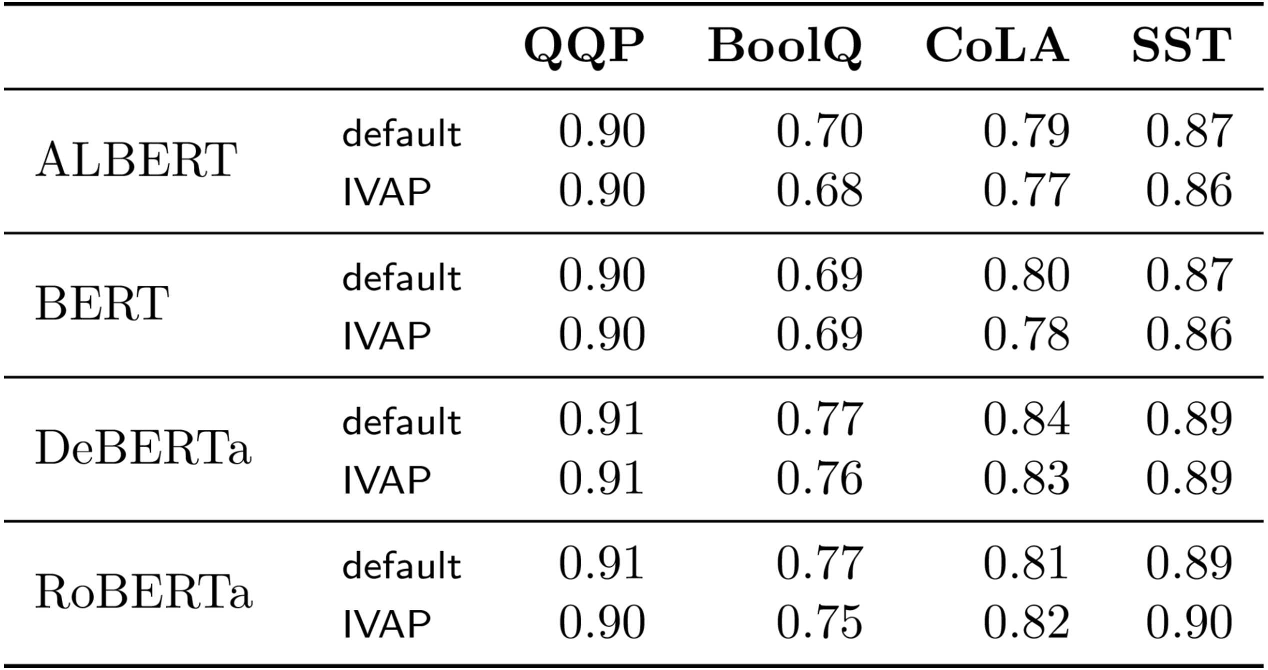
Sharpness: DeBERTa on SST



Default

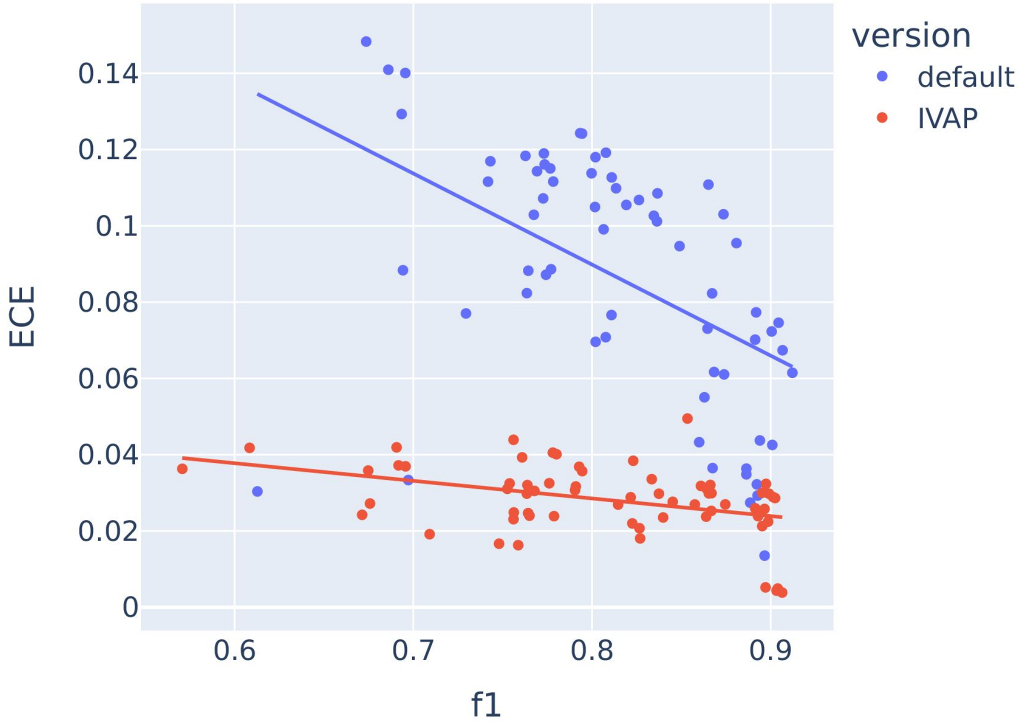


Venn–ABERS



		QQP	BoolQ	CoLA	SST
ALBERT	default	0.90	0.70	0.79	0.87
	IVAP	0.90	0.68	0.77	0.86
BERT	default	0.90	0.69	0.80	0.87
	IVAP	0.90	0.69	0.78	0.86
DeBERTa	default	0.91	0.77	0.84	0.89
	IVAP	0.91	0.76	0.83	0.89
RoBERTa	default	0.91	0.77	0.81	0.89
	IVAP	0.90	0.75	0.82	0.90

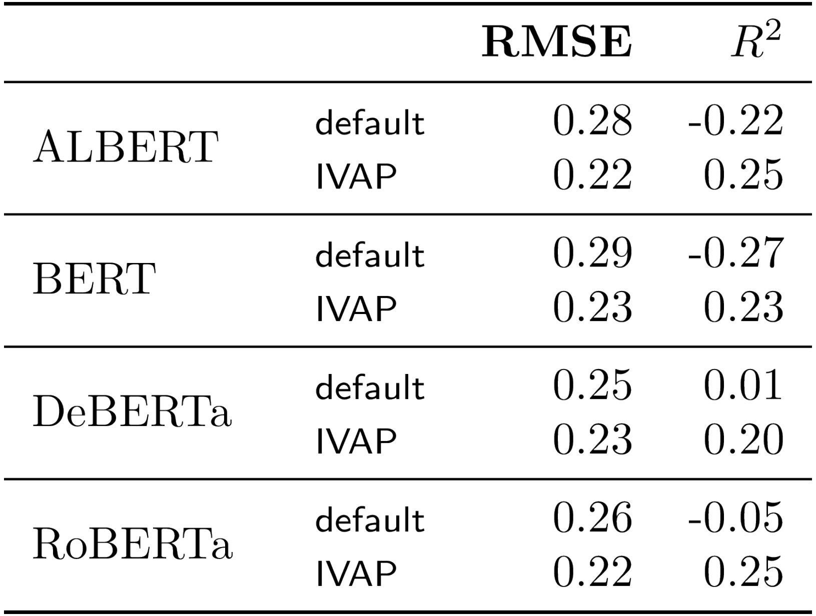
Classification performance: F1 scores for default and IVAP models.



Trend of expected calibration error versus F_1 score for all models and datasets.

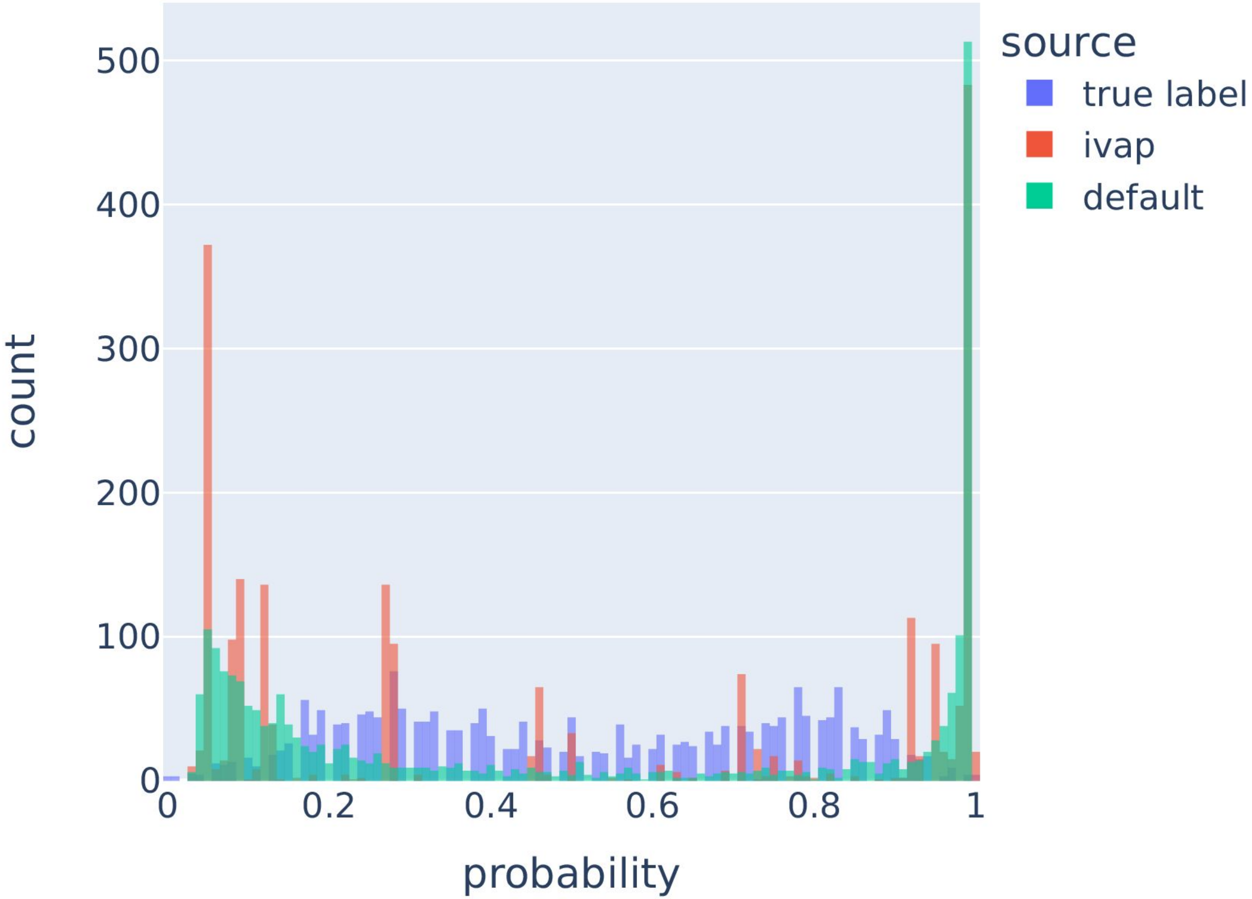
Sentiment reconstruction

- Some task are less “binary” than others
- SST originally had real numbers as labels
- Our models were trained on the “binarized” dataset
- Can we reconstruct the **degree** of positive sentiment from binary labels alone?



		RMSE	R^2
ALBERT	default	0.28	-0.22
	IVAP	0.22	0.25
BERT	default	0.29	-0.27
	IVAP	0.23	0.23
DeBERTa	default	0.25	0.01
	IVAP	0.23	0.20
RoBERTa	default	0.26	-0.05
	IVAP	0.22	0.25

Well-calibrated predictions are more aligned to human judgement



Discussion + Future Work

- Venn–ABERS predictors can be successfully applied to transformer models to obtain well-calibrated NLU predictions
  - Especially on a large dataset
  - Calibrated models retained the predictive power of the originals
- Venn–ABERS appeared to be sharper

- Extend to multiclass case
- Compare to other calibration techniques (temperature scaling)

Conclusion

- The need for reliable NLU models will continue to grow as cutting-edge research is turned into products for large audiences
  - users need to know when to trust a certain output
- A system with the ability of assessing its own uncertainty will always feel more “intelligent” than a blindly overconfident one
- This reinforces the need for accurate calibration on the path towards a better AI

Code available at https://github.com/patpizio/vennabers-for-nlu