A Transformer Conformal Predictor for Paraphrase Detection

Patrizio Giovannotti | Centre for Reliable Machine Learning, Royal Holloway University of London

Paraphrase Detection

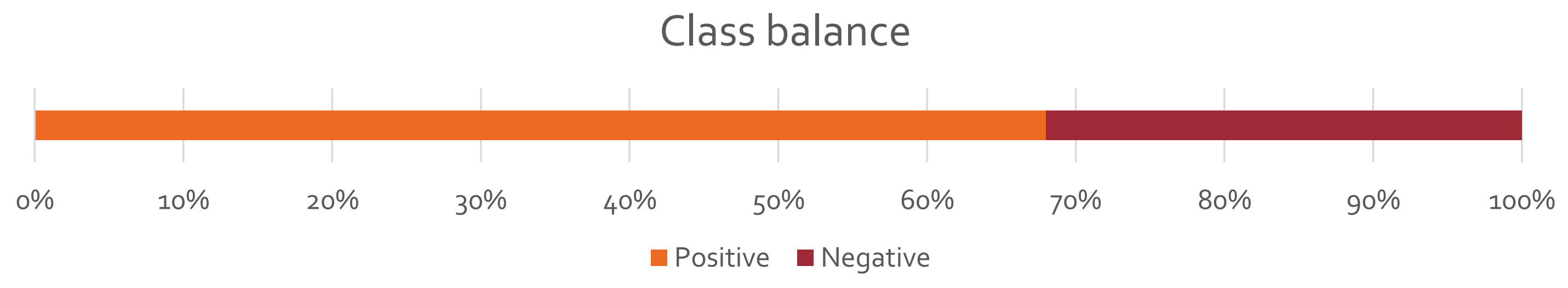
Recognise semantically equivalent sentences

Can we do it reliably?

Data

Microsoft Research Paraphrase Corpus (MRPC)

- 5801 examples (x_i, y_i)
- Each object is a sentence pair: $x_i = (s_1, s_2)$
- $y_i = 1$ if s_1 is a paraphrase of s_2, else $y_i = 0$



Each x_i can be represented by the textual concatenation $s = s_1 + s_2$

State of the Art

- **Transformers:** attention-based neural networks
- Can be pre-trained on unlabelled text
- We can fine-tune the pre-trained model on our dataset

- BERT is a popular pre-trained language model
- **DistilBERT** is a lighter version of BERT obtained via *knowledge distillation*
- We use DistilBERT as underlying algorithm

Nonconformity Measures (NCM)

For a label y:

- ***Prob:*** if $\boldsymbol{o}$ is the output layer of DistilBERT, we take
$$-\sigma(\boldsymbol{o})_y$$
as NCM, where σ is the *softmax* function

- ***Ave:*** We can extract an embedding $\boldsymbol{v} \in \mathbb{R}^{768}$ for each s from the last layer of the transformer.
If $\overline{\boldsymbol{v}}_y$ is the average embedding of the examples labelled as y, we take the Euclidean distance
$$d(\boldsymbol{v}, \overline{\boldsymbol{v}}_y)$$
as NCM for example $\boldsymbol{v}$

Predictive Performance

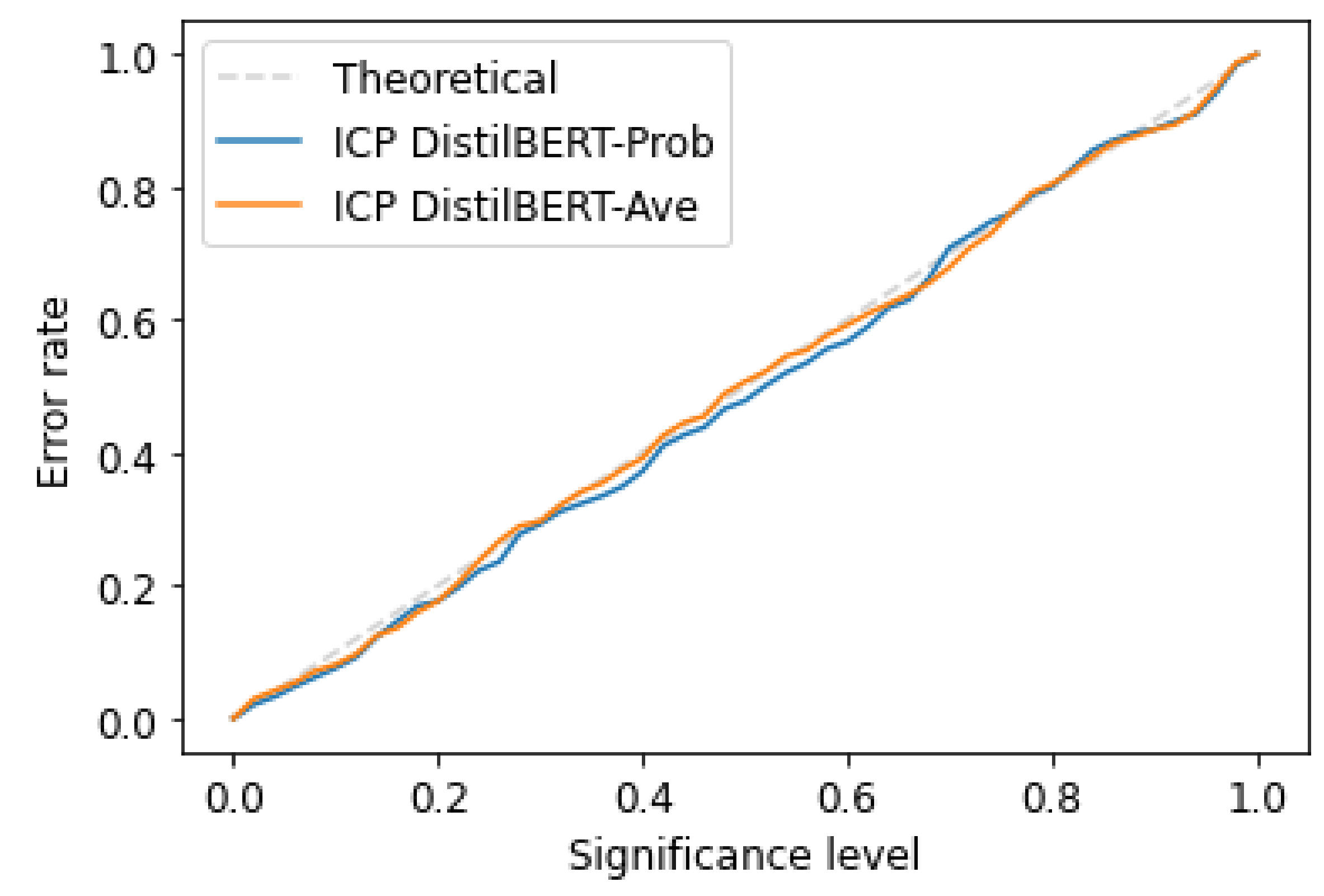
We train an Inductive Conformal Predictor (ICP) based on DistilBERT and the 2 defined NCMs

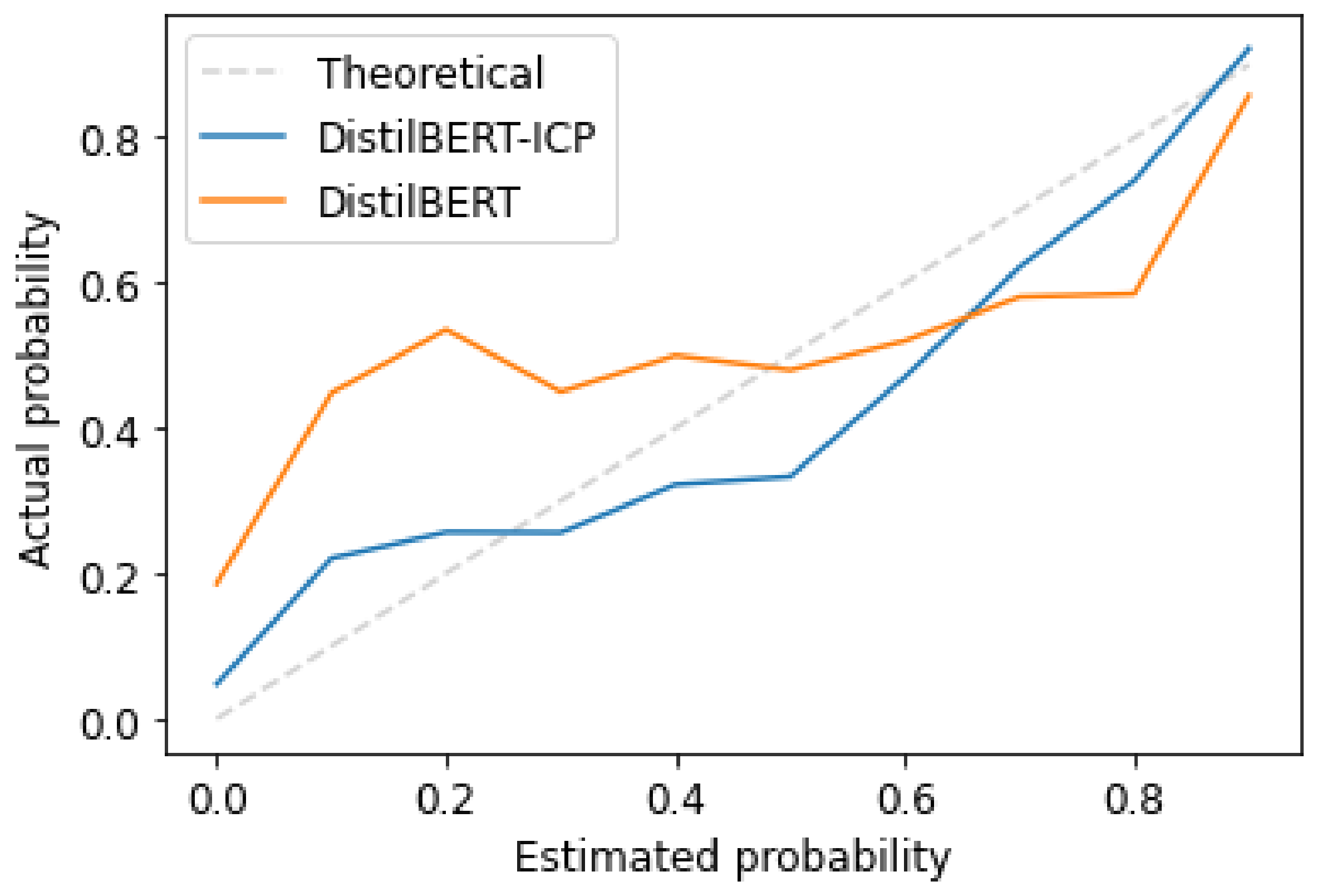
Epochs: 5 Learning rate: 5^{-5} Optimizer: Adam

Model	Acc	F_1	OF
DistilBERT	0.80	0.85	-
DistilBERT – Prob	0.80	0.84	0.13
DistilBERT – Ave	0.79	0.83	0.16

OF is Observed Fuzziness: average of the p-values for all postulated labels, except for the true labels

Validity & Calibration





Discussion

Little loss in accuracy when adding ICP, while predictions are better calibrated. Calibration still imperfect due to small size of calibration set.
In future we will try different NCMs and larger models, applied to other GLUE tasks.

References

- Sanh et al. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. ArXiv, abs/1910.01108, 2019.
- Paisios et al. A deep neural network conformal predictor for multi-label text classification. In Proceedings of the Eighth Symposium on Conformal and Probabilistic Prediction and Applications, 2019.