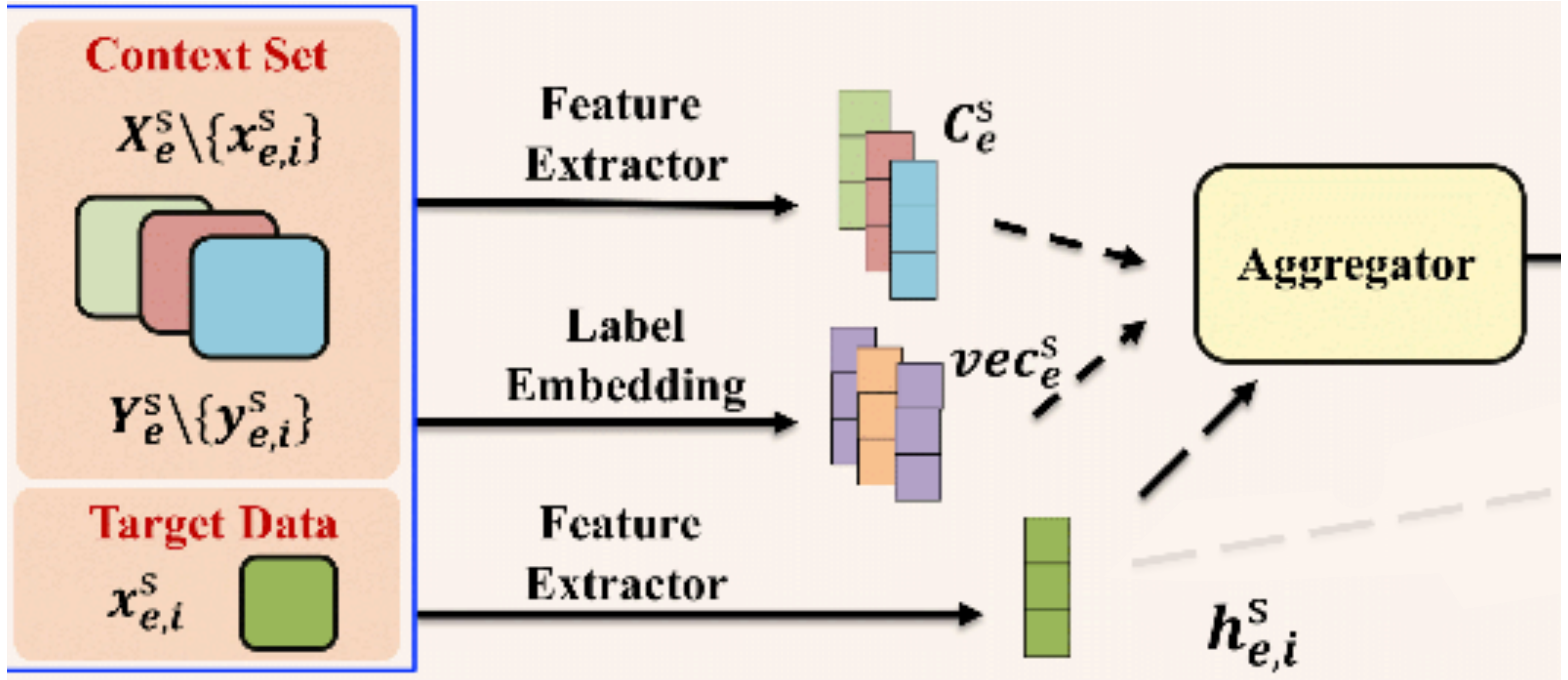
Methodology

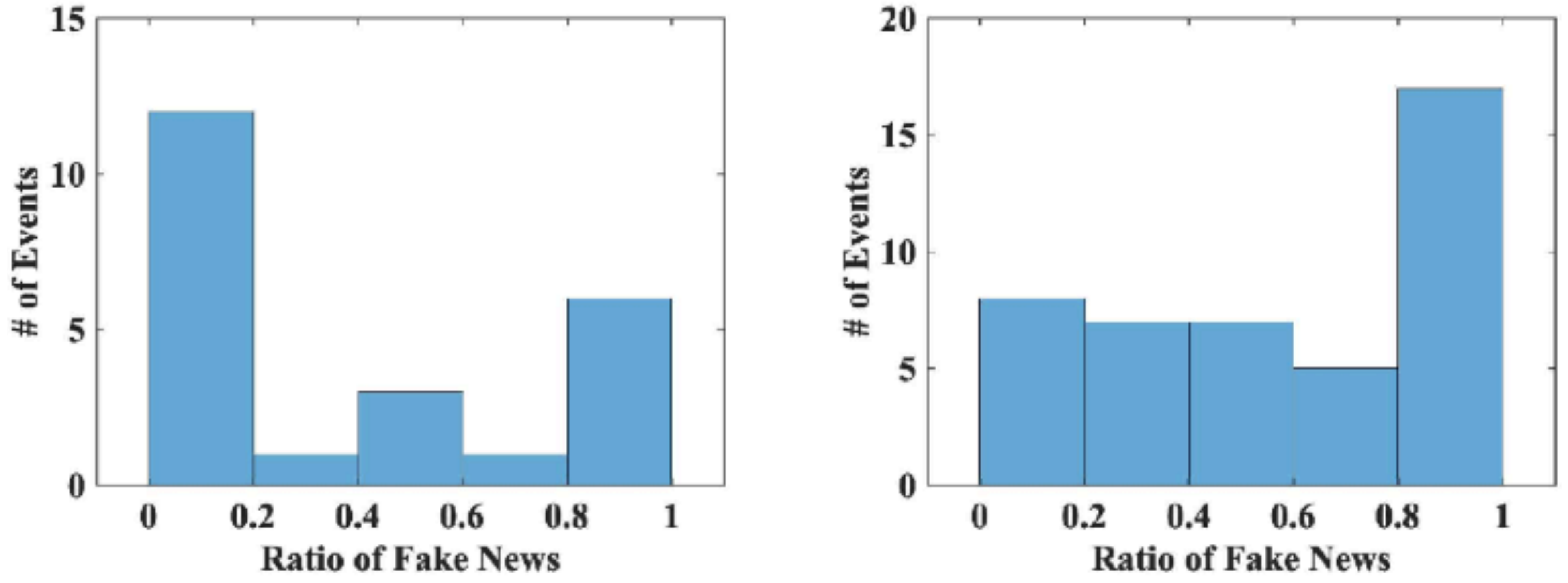
Aggregator: Limitation of Soft-attention



- $$a_i = \text{softmax}\left(\frac{Q_i K^T}{\sqrt{d}}\right)$$

- The attention mechanism with soft weight values is categorized into soft-attention.

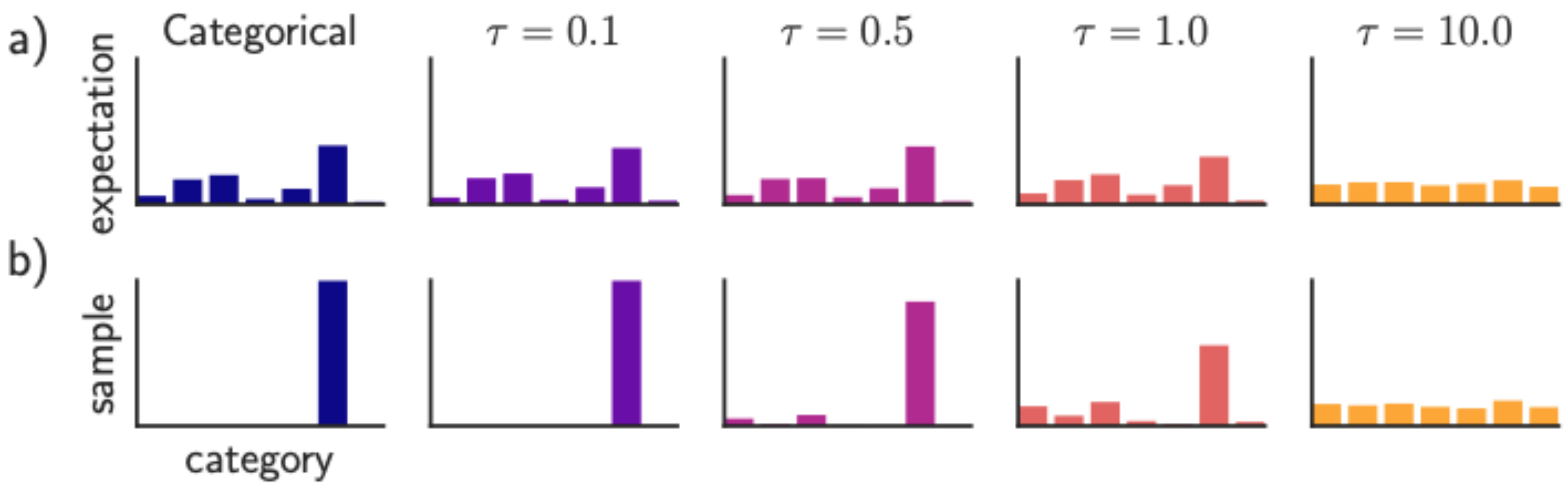
- However, soft-attention cannot effectively trim irrelevant data especially when have a context set with an imbalanced class distribution as mentioned before.



The number of events with respect to different percentages of fake news.

Methodology

Aggregator: Hard-Attention



https://arxiv.org/abs/1611.01144

- To overcome this limitation, propose to select the most related context data point instead of weighted average.

- To enable argmax operation to be differentiable, use Straight-Through (ST) Gumbel SoftMax (ICLR'17) for discretely sampling the context information given target data.

- Through gumbel-softmax, the hard-attention is able to trim the irrelevant data and draw the most informative sample for given target sample $x_{e,i}$.

- The selected data point $\mathbf{c}_{e,k} \oplus \mathbf{v}_{e,k}$ is fed into fully connected layer that top of the aggregator to adjust dimension and output context embedding $\mathbf{r}_{e,i}$.