Stylometric and Neural Features Combined Deep Bayesian Classifier for Authorship Verification

Yitao Sun, Svetlana Afanaseva and Kailash Patil

ysun@pindrop.com, safanaseva@pindrop.com

Outline

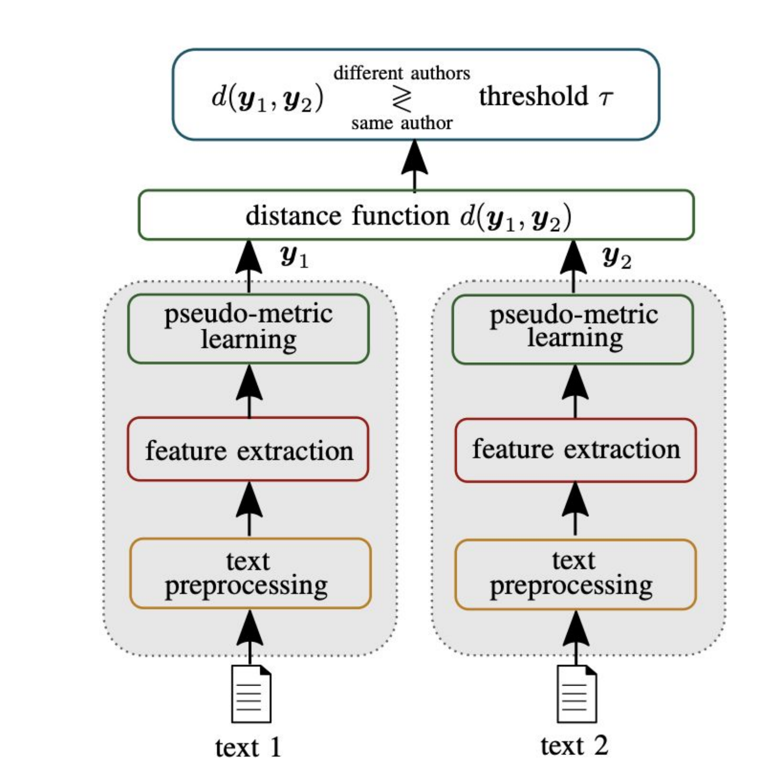
1. Idea
2. Model overview
3. Results
4. On-going work
5. Q&A

Ideas

1) Siamese Network Adhominem system (B. Boenninghoff, 2019) lacks stylometric feature
2) To improve the Adhominem system:
    a) Add Stylometric features (J. Weerasinghe, 2020), e.g frequency of function words, vocab richness, to enrich input data
    b) Use probabilistic linear discriminant analysis (PLDA) as metric layer

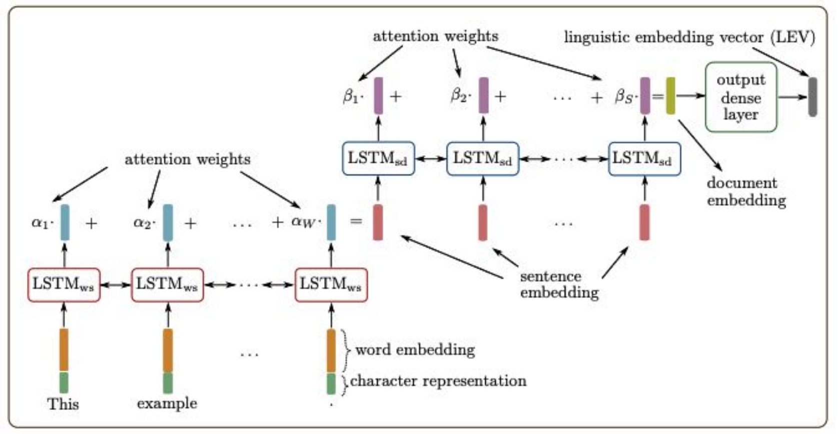
Model Overview

Siamese Network

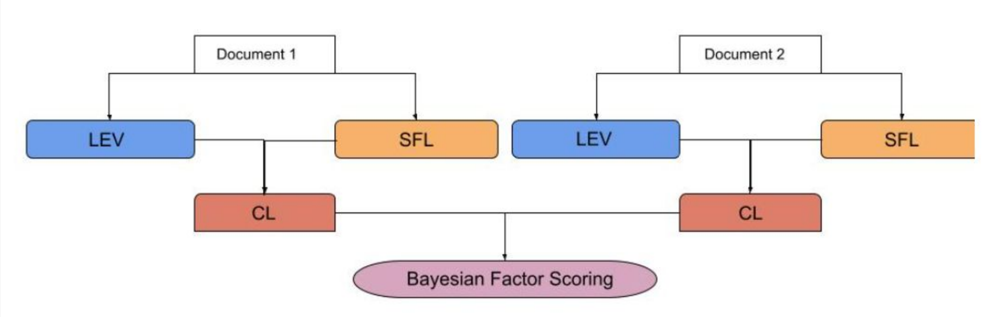


Model Overview

LEV (linguistic embedding vectors) in detail



SFL (Stylometric Feature Layers)



Bayesian Factor Scoring(B. Boenninghoff 2020)

$$\underbrace{\boldsymbol{y}}_{\text{combined layers}} = \underbrace{\boldsymbol{x}}_{\text{author's writing style}} + \underbrace{\boldsymbol{\epsilon}}_{\text{noise term}}$$

- $\mathcal{H}_s$: The two documents were written by the same person,
- $\mathcal{H}_d$: The two documents were written by two different persons.

Same-author pair probability

$$p\left(\boldsymbol{y}_1, \boldsymbol{y}_2 \mid \mathcal{H}_s\right) = \frac{p\left(\boldsymbol{y}_1, \boldsymbol{y}_2 \mid \boldsymbol{x}_0, \mathcal{H}_s\right) p\left(\boldsymbol{x}_0 \mid \mathcal{H}_s\right)}{p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}_1, \boldsymbol{y}_2, \mathcal{H}_s\right)} = \frac{p\left(\boldsymbol{y}_1 \mid \boldsymbol{x}_0\right) p\left(\boldsymbol{y}_2 \mid \boldsymbol{x}_0\right) p\left(\boldsymbol{x}_0\right)}{p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}_1, \boldsymbol{y}_2\right)}$$

Different-author pair probability

$$p\left(\boldsymbol{y}_1, \boldsymbol{y}_2 \mid \mathcal{H}_d\right) = p\left(\boldsymbol{y}_1 \mid \mathcal{H}_d\right) p\left(\boldsymbol{y}_2 \mid \mathcal{H}_d\right) = \frac{p\left(\boldsymbol{y}_1 \mid \boldsymbol{x}_1\right) p\left(\boldsymbol{x}_1\right)}{p\left(\boldsymbol{x}_1 \mid \boldsymbol{y}_1\right)} \cdot \frac{p\left(\boldsymbol{y}_2 \mid \boldsymbol{x}_2\right) p\left(\boldsymbol{x}_2\right)}{p\left(\boldsymbol{x}_2 \mid \boldsymbol{y}_2\right)}$$

Bayesian Factor Scoring

Final bayesian score generation

$$
\begin{aligned}
\text{score}(y_1, y_2) &= \log p(y_1, y_2 \mid \mathcal{H}_s) - \log p(y_1, y_2 \mid \mathcal{H}_d) \\
&= \log p(\boldsymbol{x}_0) - \log p(\boldsymbol{x}_1) - \log p(\boldsymbol{x}_2) \\
&\quad + \log p(y_1 \mid \boldsymbol{x}_0) + \log p(y_2 \mid \boldsymbol{x}_0) - \log p(y_1 \mid \boldsymbol{x}_1) - \log p(y_2 \mid \boldsymbol{x}_2) \\
&\quad - \log p(\boldsymbol{x}_0 \mid y_1, y_2) + \log p(\boldsymbol{x}_1 \mid y_1) + \log p(\boldsymbol{x}_2 \mid y_2)
\end{aligned}
$$

Bayesian Factor Scoring

To learn the probability layer from the data

$$p(\mathcal{H}_s|\boldsymbol{y}_1, \boldsymbol{y}_2) = \frac{p(\boldsymbol{y}_1, \boldsymbol{y}_2|\mathcal{H}_s)}{p(\boldsymbol{y}_1, \boldsymbol{y}_2|\mathcal{H}_s) + p(\boldsymbol{y}_1, \boldsymbol{y}_2|\mathcal{H}_d)} = \mathrm{Sigmoid}\big(\mathrm{score}(\boldsymbol{y}_1, \boldsymbol{y}_2)\big)$$

The loss function

$$\mathcal{L}_\phi = l \cdot \log\left\{p(\mathcal{H}_s|\boldsymbol{y}_1, \boldsymbol{y}_2)\right\} + (1 - l) \cdot \log\left\{1 - p(\mathcal{H}_s|\boldsymbol{y}_1, \boldsymbol{y}_2)\right\}$$

Results

Test Results of PAN 2023 Training Dataset

Model	*AUC*	*C@1*	$f_0 5_u$	*F1*	*brier*	*overall*
Naive, Distance-based	0.493	0.497	0.553	0.664	0.741	0.589
Method-based text compression	0.504	0.033	0.048	0.621	0.75	0.391
DML without SFL	0.503	0.523	0.492	0.357	0.603	0.495
UAL without SFL	0.499	0.52	0.477	0.336	0.593	0.485
BFS without SFL	0.47	0.502	0.474	0.37	0.597	0.483
DML with SFL	0.523	0.499	0.605	0.522	0.73	0.576
UAL with SFL	0.568	0.492	0.584	0.467	0.747	0.571
BFS with SFL	**0.658**	**0.662**	**0.739**	**0.735**	**0.762**	**0.711**

On-going work and thoughts

1. Didn’t finetune the threshold, can do better next year
2. Might consider EER(Equal Error Rate) as one of the metric
3. Does online text(word) distribute the same way as transcribed text
4. Use other popular pretrained transformer with BFS(Bayesian Factor Scoring) to boost performance

Q&A