Feature Vector Difference based Authorship Verification for Open-World Settings

PAN at CLEF 2021

Janith Weerasinghe
janith@nyu.edu

Rhia Singh
rhia.singh@macaulay.cuny.edu

Rachel Greenstadt
greenstadt@nyu.edu

NYU TANDON SCHOOL OF ENGINEERING

MACAULAY HONORS COLLEGE

PAN 2021 Authorship Verification Task

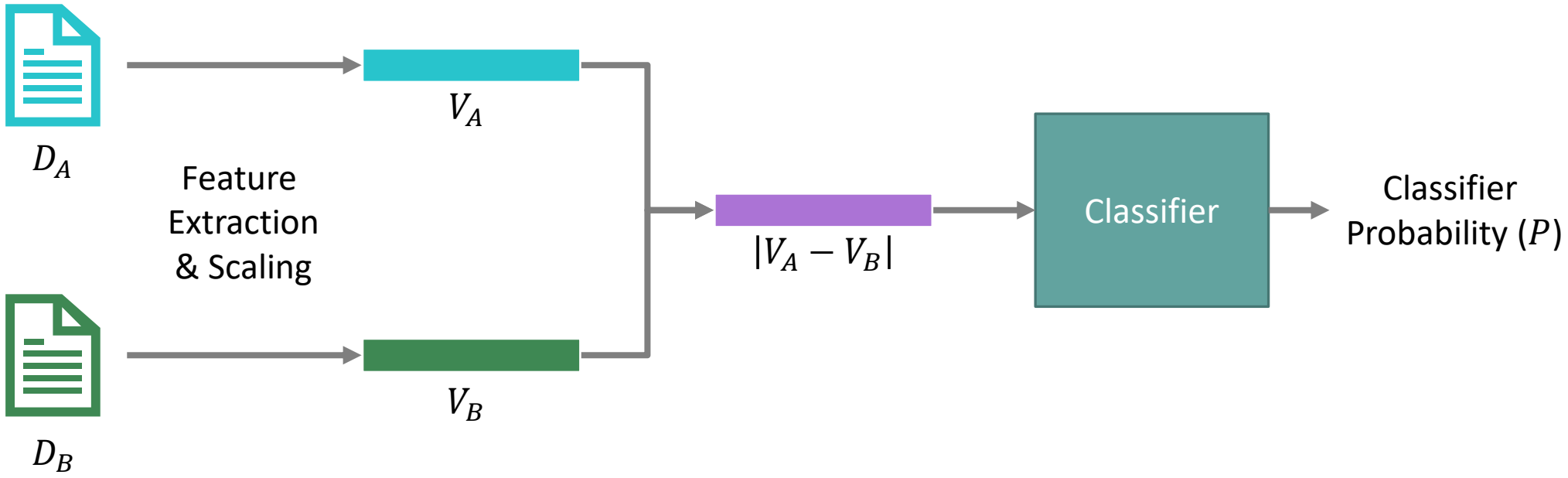
- Predict if two documents are written by the same author. The documents are written by previously unseen authors on previously unseen topics

```
1.{"id": "6cced668-6e51-5212-873c-717f2bc91ce6", "fandoms": ["Fandom 1", "Fandom 2"], "pair": ["Text 1...", "Text 2..."]}
2.{"id": "ae9297e9-2ae5-5e3f-a2ab-ef7c322f2647", "fandoms": ["Fandom 3", "Fandom 4"], "pair": ["Text 3...", "Text 4..."]}
```

```
1.{"id": "6cced668-6e51-5212-873c-717f2bc91ce6", "same": true, "authors": ["1446633", "1446633"]}
2.{"id": "ae9297e9-2ae5-5e3f-a2ab-ef7c322f2647", "same": false, "authors": ["1535385", "1998978"]}
```

- Training Dataset:
  - Small: 52,590 Document pairs
  - Large: 275,486 Document pairs
  - ~ 21,000 characters, ~4,800 tokens per document

Approach


D_A
D_B
Feature Extraction & Scaling
V_A
V_B
$|V_A - V_B|$
Classifier
Classifier Probability (P)

Features

- Character tri-grams (TF – IDF)
- Special Characters (TF – IDF)
- Frequency of Function Words
- Average number of characters per word
- Distribution of word-lengths (1-10)
- Vocabulary Richness measures*
- Unique Spellings (fraction of tokens)*
  - Commonly misspelled words, British spelling of words, and popular online abbreviations

* New in this year's model

Features

Example:
```
The Soviets had already been merciless, ruthless
as the next army.
```

POS Tags:
```
[('The', 'DT'), ('Soviets', 'NNPS'), ('had', 'VBD'),
('already', 'RB'), ('been', 'VBN'), ('merciless',
'NN'), (',', ','), ('ruthless', 'NN'), ('as', 'IN'),
('the', 'DT'), ('next', 'JJ'), ('army', 'NNP'), ('.',
'.')]
```

Parse Tree:
```
(S
      (NP The/DT Soviets/NNPS)
      (VP had/VBD already/RB been/VBN)
      (NP merciless/NN)
      ,/,
      (NP ruthless/NN)
      as/IN
      (NP the/DT next/JJ army/NNP)
)
```

- POS-Tag tri-grams (TF – IDF)
- POS-Tag chunk tri-grams (TF – IDF) :
  - `[NP VP NP , NP IN NP .]`
- POS tag chunk construction (TF – IDF) :
  - `NP[DT NNPS], VP[VBD RB VBN], NP[NN], NP[NN], NP[DT JJ NNP]`
- Function-word and POS tag hybrid tri-grams*:
  - `[The NNPS had already been NN , NN as the JJ NNP]`
- POS tag ratios*

* New in this year's model

Classifier

- Logistic Regression Classifier
  - Trained using Stochastic Gradient Descent

- Trained using 75% data
  - No training set authors in the test set

Results

Dataset / Model Version	AUC	C@1	F0.5U	F1-Score	Brier
Small Dataset, early submission	0.955	0.890	0.894	0.889	0.919
Small Dataset, local test set	0.965	0.903	0.928	0.903	0.925
Small Dataset (trained on full dataset), final evaluation	**0.967**	**0.910**	**0.907**	**0.927**	**0.929**
Large Dataset, local test set	0.967	0.909	0.918	0.915	0.928
Large Dataset, final evaluation	**0.972**	**0.917**	**0.916**	**0.926**	**0.934**

Thank You!

Questions:
Janith Weerasinghe: janith@nyu.edu

Source Code and Models:
https://github.com/janithnw/pan2021_authorship_verification

Acknowledgements:
PAN 2021 organizers and reviewers
Funded by NSF Grant 1931005 and the McNulty Foundation