Feature Vector Difference based Neural Network and Logistic Regression Models for Authorship Verification

PAN at CLEF 2020

Janith Weerasinghe
janith@nyu.edu

Rachel Greenstadt
greenstadt@nyu.edu

NYU TANDON SCHOOL OF ENGINEERING

Image from: http://www.wallpaperaccess.com

PAN 2020 Authorship Verification Task

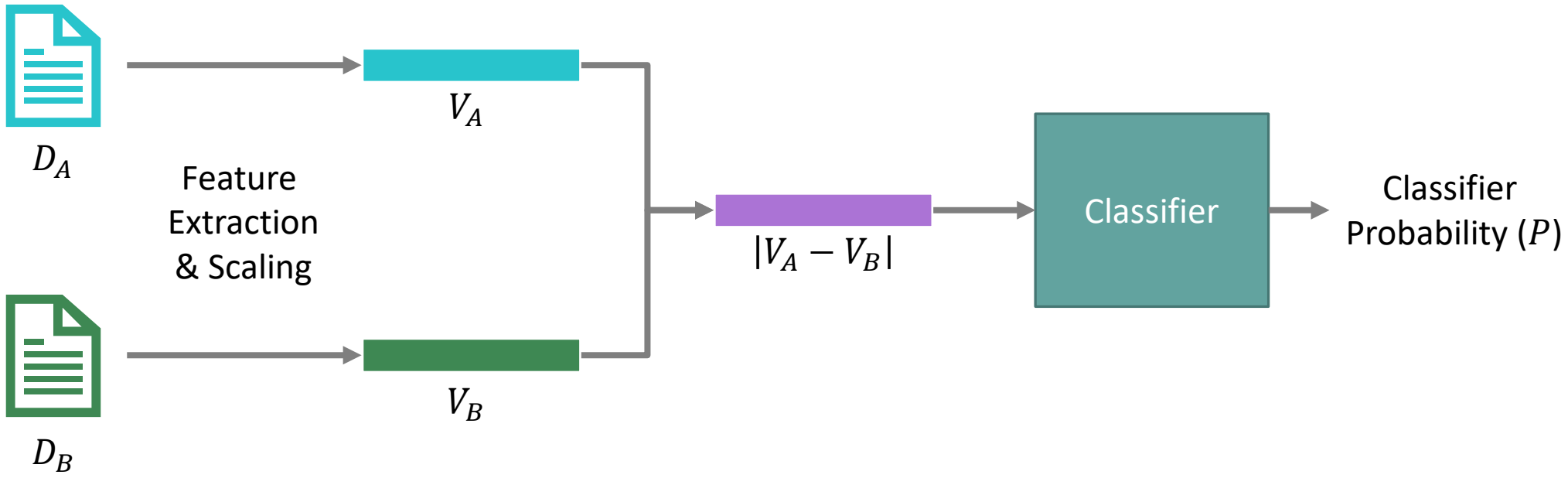
- "…deciding whether two texts have been written by the same author based on comparing the texts' writing styles"

```
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 n-grams (TF – IDF)
- Special Characters (TF – IDF)
- Frequency of Function Words
- Number of characters
- Number of words
- Average number of characters per word
- Distribution of word-lengths (1-10)
- Vocabulary Richness: The ratio of hapax-legomenon and dis-legomenon

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 .]`
- NP and VP construction (TF – IDF) :
  - `NP[DT NNPS], VP[VBD RB VBN], NP[NN], NP[NN], NP[DT JJ NNP]`

Classifiers

- Logistic Regression Classifier
  - Trained on smaller dataset
- Neural Network
  - Single hidden layer
  - Trained on larger dataset
- Other classifier experiments: SVM and Random Forest

Results

Dataset / Model	AUC	C@1	F0.5U	F1-Score
Small, Logistic Regression	0.939	0.833	0.817	0.860
Large, Neural Network	0.953	0.880	0.882	0.891

Future Work

- Feature analysis and misclassification analysis
- Optimize for C@1
- Test on unseen authors
- Reduce document size

Thank You!

Questions:

Janith Weerasinghe: janith@nyu.edu

Source Code and Models:

https://github.com/janithnw/pan2020_authorship_verification

Acknowledgements:

PAN 2020 organizers and reviewers

Funded by NSF Grant 1931005

My favorite fanfiction from the training set:

Id: '32c367de-1a55-542f-8efb-a65b6b68a0e6'

Pair[1]: 'Pitch laughed darkly, "As for the fact I have you all right where I want you. Now face the newest immortal amongst us who has joined together with me to defeat the whole lot of you". Seconds went by without a sound but that quickly ended as Tooth pointed to the sky where a rainbow was coming down like a falling star. Then it opened up its mouth. "NYAN … … …