Authorship Verification with neural networks via stylometric feature concatenation

Authorship Verification. PAN at CLEF 2021
22 -23 September 2021

Name
Antonio Menta Garuz (amenta@invi.uned.es) NLP & IR Research Group (UNED), Spain
Ana Garcia Serrano (agarcia@lsi.uned.es) NLP & IR Research Group (UNED), Spain

Our Approach

Features: Based on Stylometric feature extraction process.

- Character-level n-grams.
- Punctuation marks.

Model input:

- feature vector difference between texts as input for NN.

Model: Binary classifier based on neural networks.

- One NN for each feature.
- Concatenate output vector representation into a final NN for decision making.

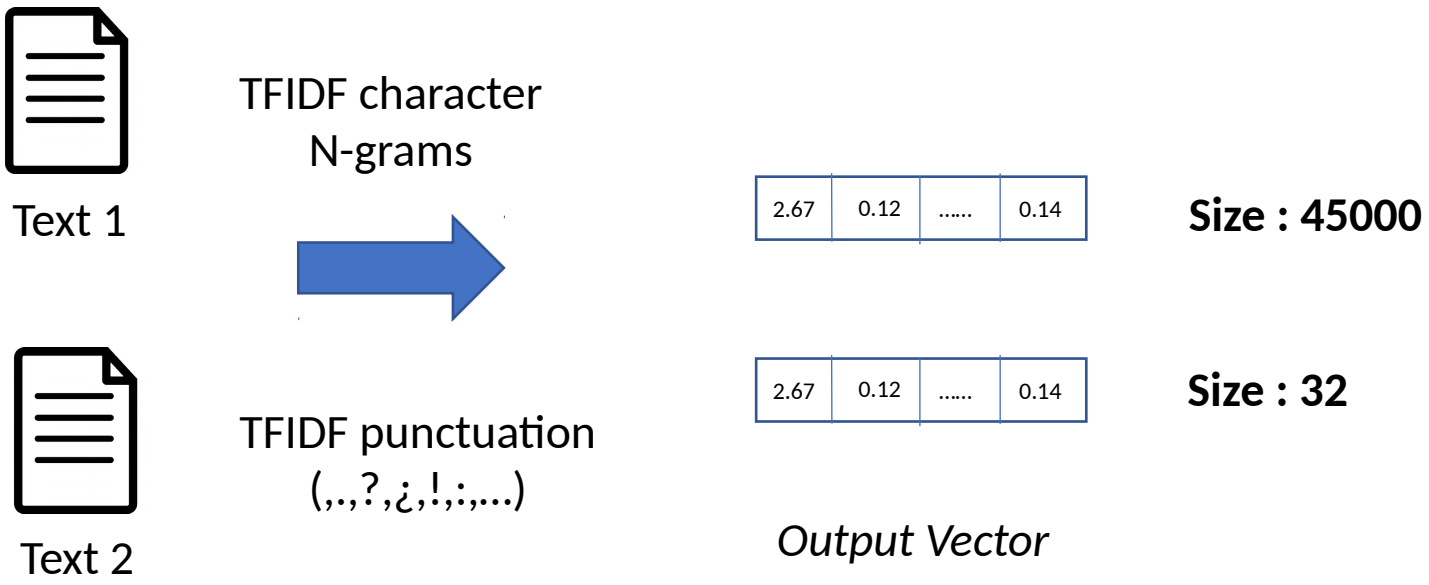
Features

Objective:

Reduce high dimensionality of character-ngrams features vs punctuation marks.

- Dimension reduction techniques: PCA. ✗
- Solution: Obtain a latent vector for each feature. ✓

Features



- *J. Weerasinghe and R. Greenstadt, "Feature Vector Difference based Neural Network and Logistic Regression Models for Authorship Verification Notebook for PAN at CLEF 2020"*

- *E. Araujo-Pino, H. Gómez-Adorno, and G. Fuentes-Pineda, "Siamese Network applied to Authorship Verification Notebook for PAN at CLEF 2020."*

Neural Network

Character-level n-grams:

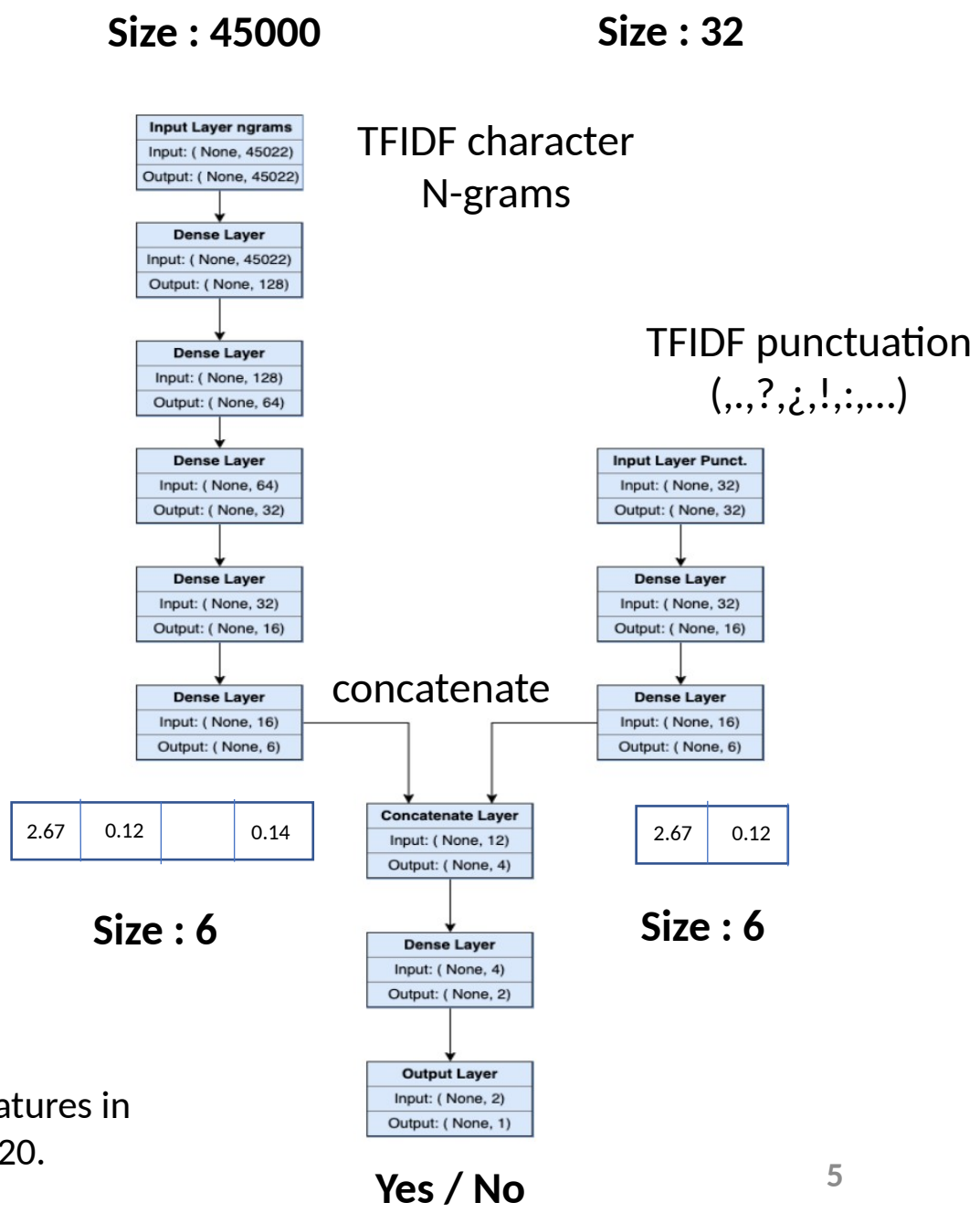
- 6 Fully Connected Layers
- From dimension 45000 to 6

Punctuation marks:

- 3 Fully Connected Layers
- From dimension 32 to 6

Concatenate layers:

- 3 Fully Connected Layers



E. Akcapinar Sezer, H. Sever, and P. Canbay, “Deep Combination of Stylometry Features in Forensic Authorship Analysis,” Int. J. Inf. Secur. Sci., vol. 9, no. 3, pp. 154–163, 2020.

Results

Results obtained at Authorship Verification shared task:

Set	AUC	c@1	F1	F0.5u	Brier	Overall
Large	0.9635	0.9024	0.8990	0.9186	0.9155	**0.9198**
Small	0.9385	0.8662	0.8620	0.8787	0.8762	**0.8843**

"**Authorship Verification with neural networks via stylometric feature concatenation**" (Menta and Garcia-Serrano, 2021)

CODE: https://github.com/Hisarlik/Authorship-verification

Conclusion

- Stylometric features can achieve competitive results.
- Neural Networks work well with both datasets. (small and large)
- The Importance of each feature can be modified by varying its output vector.

Character-level n-grams

2.67	0.12	...	0.14

Size : 6

Punctuation marks

2.67	0.12

Size : 6

Next Steps

- **Increase the number and type of features used:**
  - Lexical features: Vocabulary richness
  - Syntactic features: Part-of-speech, phrase structure
  - Structural features: Average frequencies of word length, paragraph length,…
- **Improve hyperparameters tuning of the neural network.**
  - Automated hyperparameter optimization methods

Authorship Verification with neural networks via stylometric feature concatenation

Authorship Verification. PAN at CLEF 2021
22 -23 September 2021

Name
Antonio Menta Garuz (amenta@invi.uned.es) NLP & IR Research Group (UNED), Spain
Ana Garcia Serrano (agarcia@lsi.uned.es) NLP & IR Research Group (UNED), Spain