17th International Conference on Natural Language Processing

BertAA: BERT fine-tuning for Authorship Attribution

Maël Fabien[1,2], Esaú Villatoro-Tello[1,3], Petr Motlicek[1], and Shantipriya Parida[1]

1 idiap RESEARCH INSTITUTE

2 EPFL

3 UNIVERSIDAD AUTÓNOMA METROPOLITANA Casa abierta al tiempo

ROXANNE

Outline

1. Introduction to Authorship Attribution
2. Related works
3. BertAA: Bert fine-tuning for AA
4. Authorship Attribution corpora
5. Results
6. Future Works
7. Conclusion

Authorship Analysis

Author Profiling	Authorship Attribution	Authorship Verification

Attributing a text to the correct author among of closed set of potential writers (e.g. 5, 10, 25, 50, 75 or 100 authors)

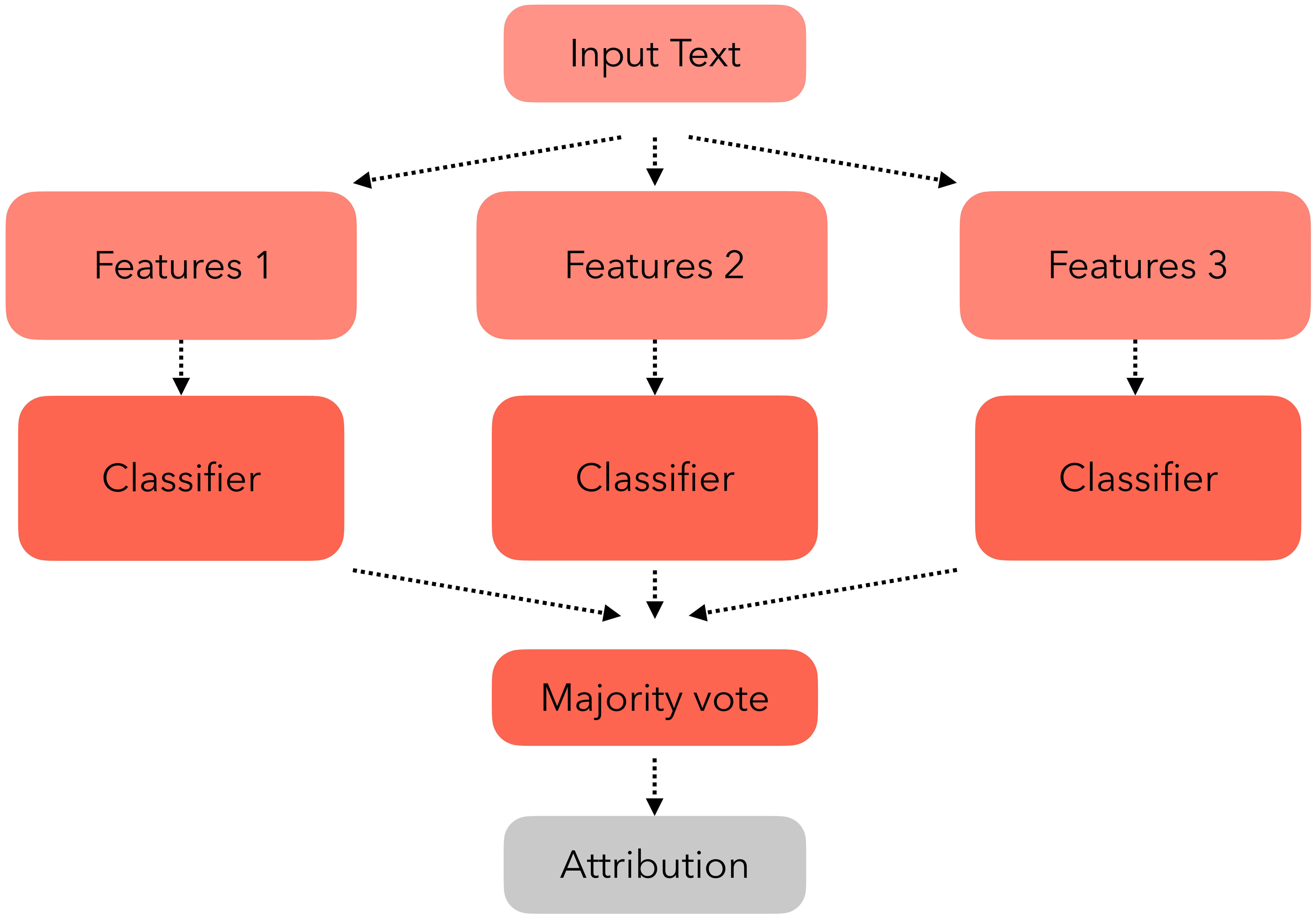
Authorship Attribution

Authorship Attribution

- Plagiarism detection
- Historical Literature
- Forensic investigations

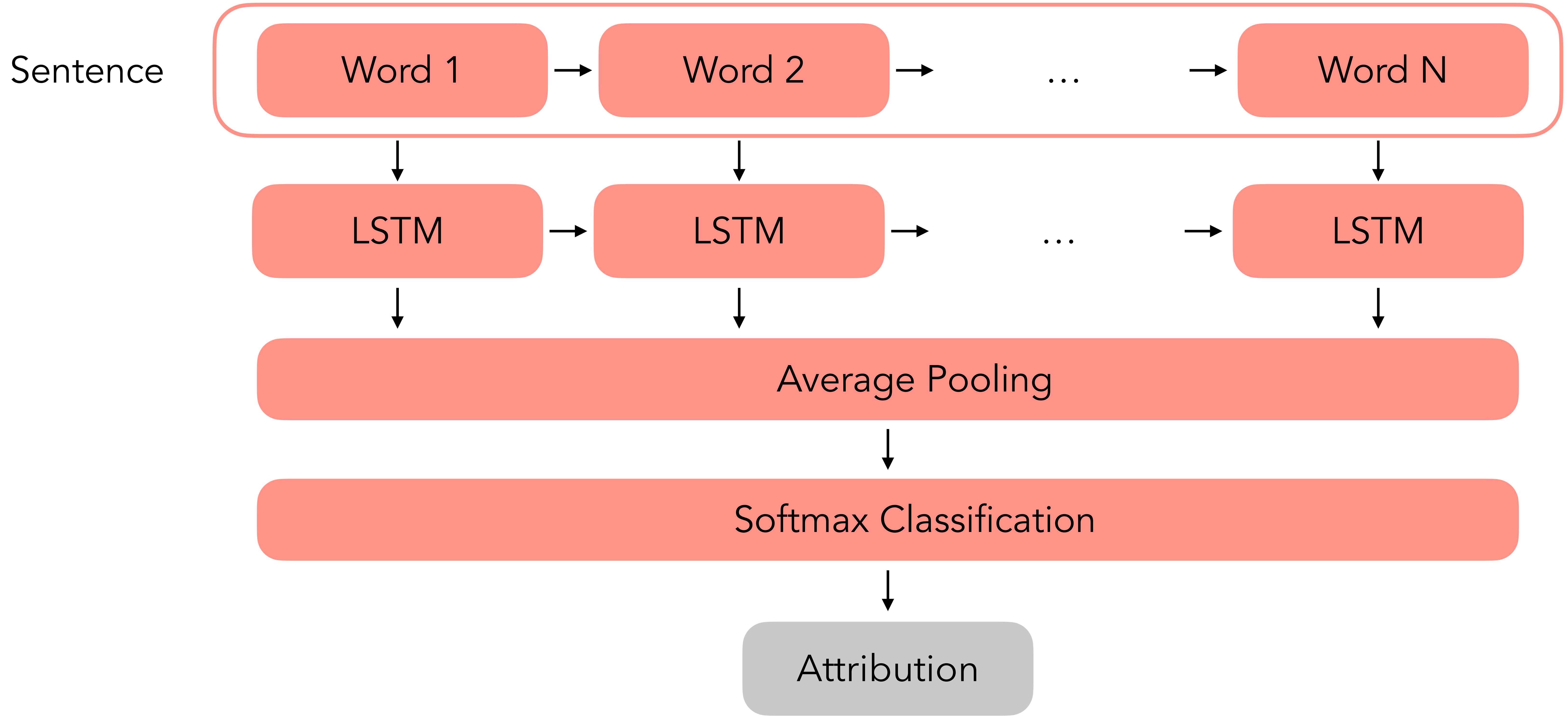
Traditional methods

Ensemble models



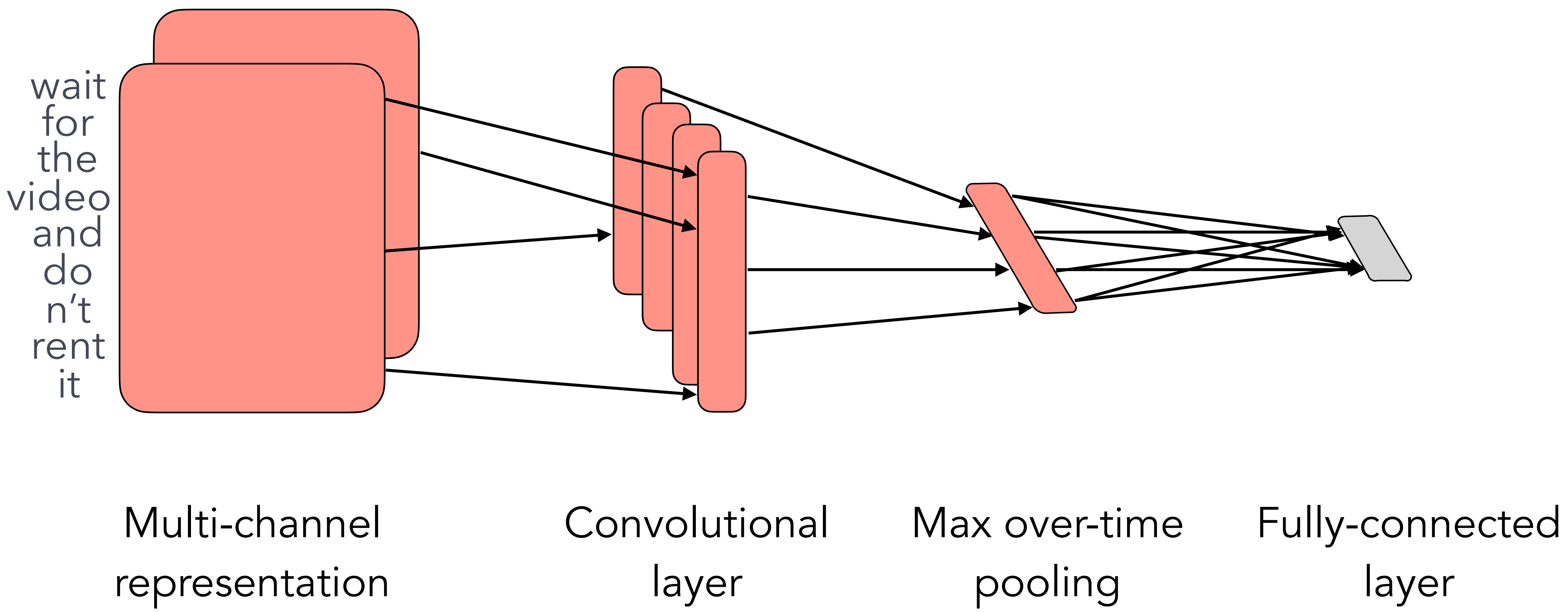
Deep-learning methods

Recurrent Neural Networks

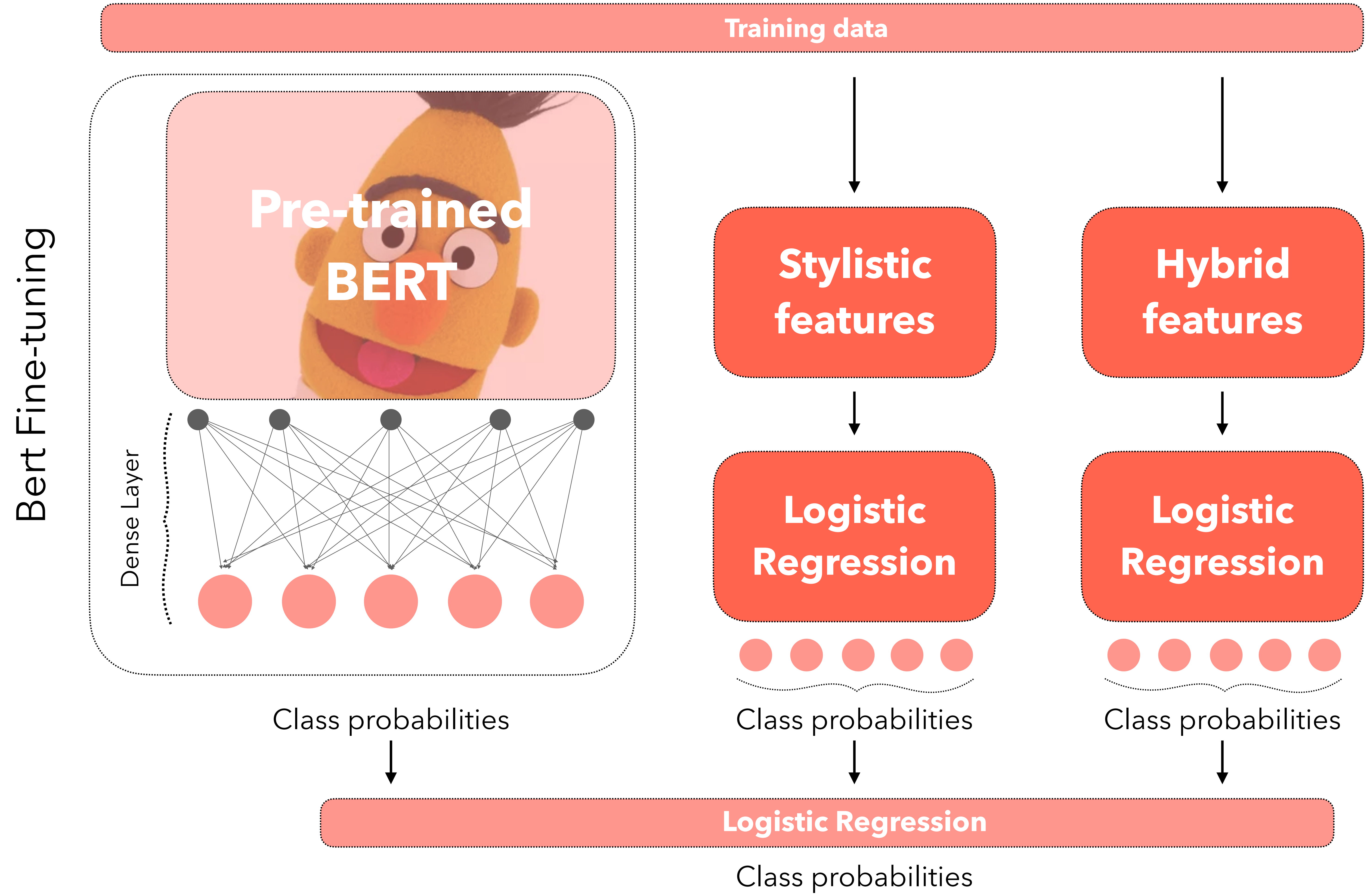


Deep-learning methods

Convolutional Neural Networks



Architecture



BertAA

+ Style

+ Hybrid

External features

Stylistic

- Length of text
- Number of words
- Average length of words
- Number of short words
- Proportion of digits and capital letters
- Individual letters and digits frequencies
- Hapax-legomena
- Frequency of 12 punctuation marks

Hybrid

- Frequency of the 100 most frequent character-level bi-grams
- Frequency of the 100 most frequent character-level tri-grams

Corpora

Dataset	Number of tokens	Number of texts
Enron	± 200	± 10’000
IMDb	± 100	± 3000
IMDb 62	340	1000
Blog	± 90	± 2500

How does the performance compare to SOTA?

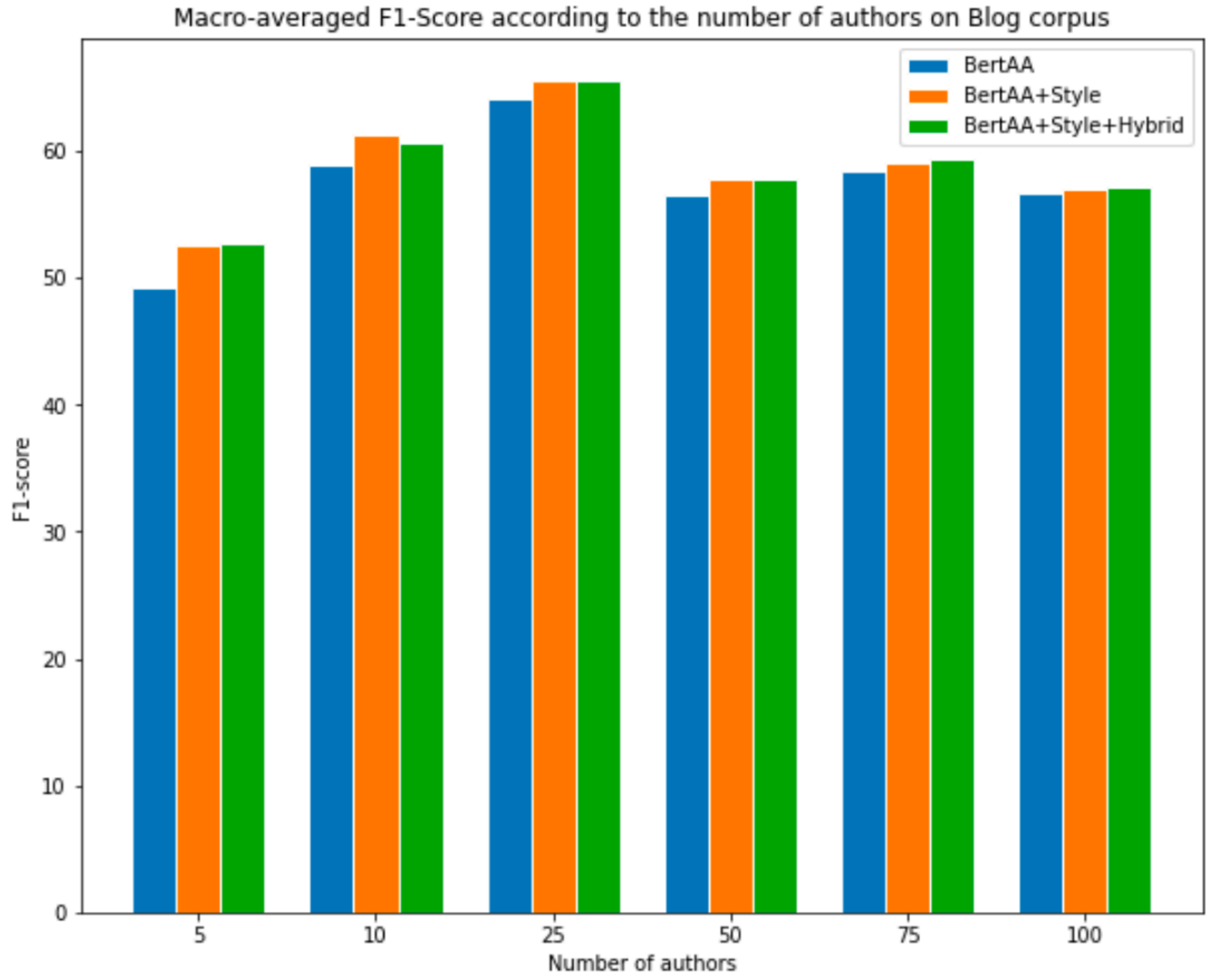
Dataset	N-Authors	Baslines			Proposed Method		
		Stylo.	Char N-gram	TF-IDF	BertAA	+ Style	+ Style + Hybrid
Enron	5	75.0	84.4	98.0	**99.95**	**99.95**	**99.95**
	10	54.9	70.5	96.4	**99.1**	**99.1**	**99.1**
	25	35.6	53.2	92.7	**98.7**	**98.7**	**98.7**
	50	20.4	44.8	90.8	98.1	**98.2**	**98.2**
	75	17.3	40.6	90.1	**97.6**	97.5	97.5
	100	15.8	36.9	88.3	97.0	97.0	**97.1**
IMDb	5	65.8	92.1	98.1	**99.6**	**99.6**	**99.6**
	10	44.6	79.2	93.9	98.1	**98.2**	**98.2**
	25	25.5	55.8	84.1	**93.2**	92.9	92.9
	50	17.4	44.2	82.1	**90.7**	90.6	90.6
	75	14.7	37.6	79.2	**88.3**	87.8	87.8
	100	11.8	33.6	76.6	**86.1**	85.3	85.4
Blog	5	34.7	40.0	45.7	**61.3**	59.7	59.8
	10	18.9	31.9	45.0	**65.4**	62.4	62.4
	25	9.9	23.4	42.0	**65.3**	64.4	64.4
	50	6.2	15.7	41.4	**59.7**	58.7	58.7
	75	5.0	15.7	42.2	**60.9**	59.0	59.2
	100	4.2	13.8	40.5	**58.8**	57.3	57.6

Approach	10	50
Impostors (Koppel and Winter, 2014)	35.4	22.6
SCAP (Frantzeskou et al., 2006)	48.6	41.6
LDAH-S (El et al.)	52.5	18.3
CNN (Ruder et al., 2016)	61.2	49.4
Continuous N-gram (Sari et al., 2017)	61.3	52.8
N-gram CNN (Zhang et al., 2018)	63.7	53.1
Syntax CNN (Zhang et al., 2018)	64.1	56.7
BertAA	**65.4**	**59.7**

Accuracy on the Blog Authorship Corpus

+**5.3**% relative improvement compared to SOTA

Are external features useful?

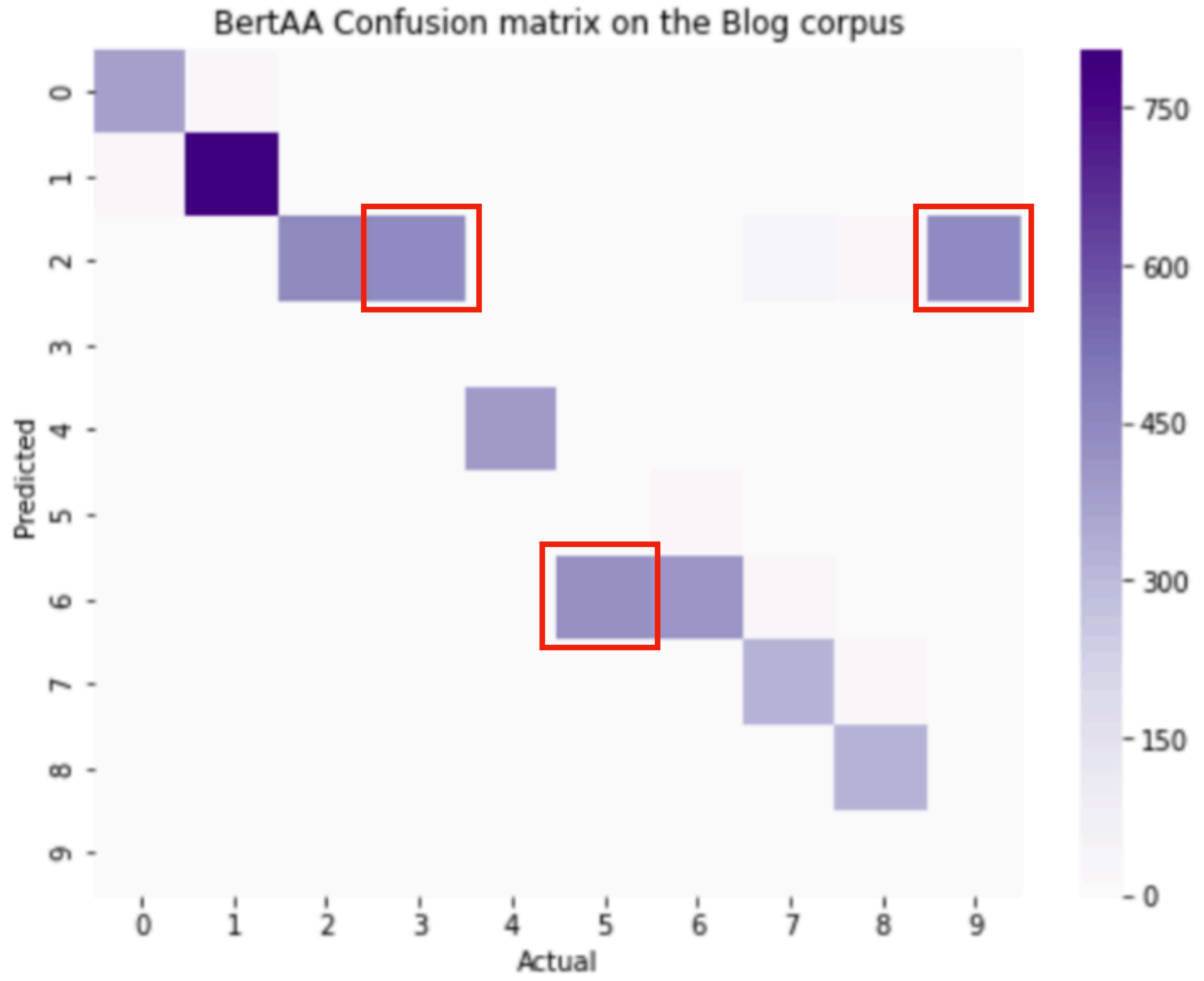


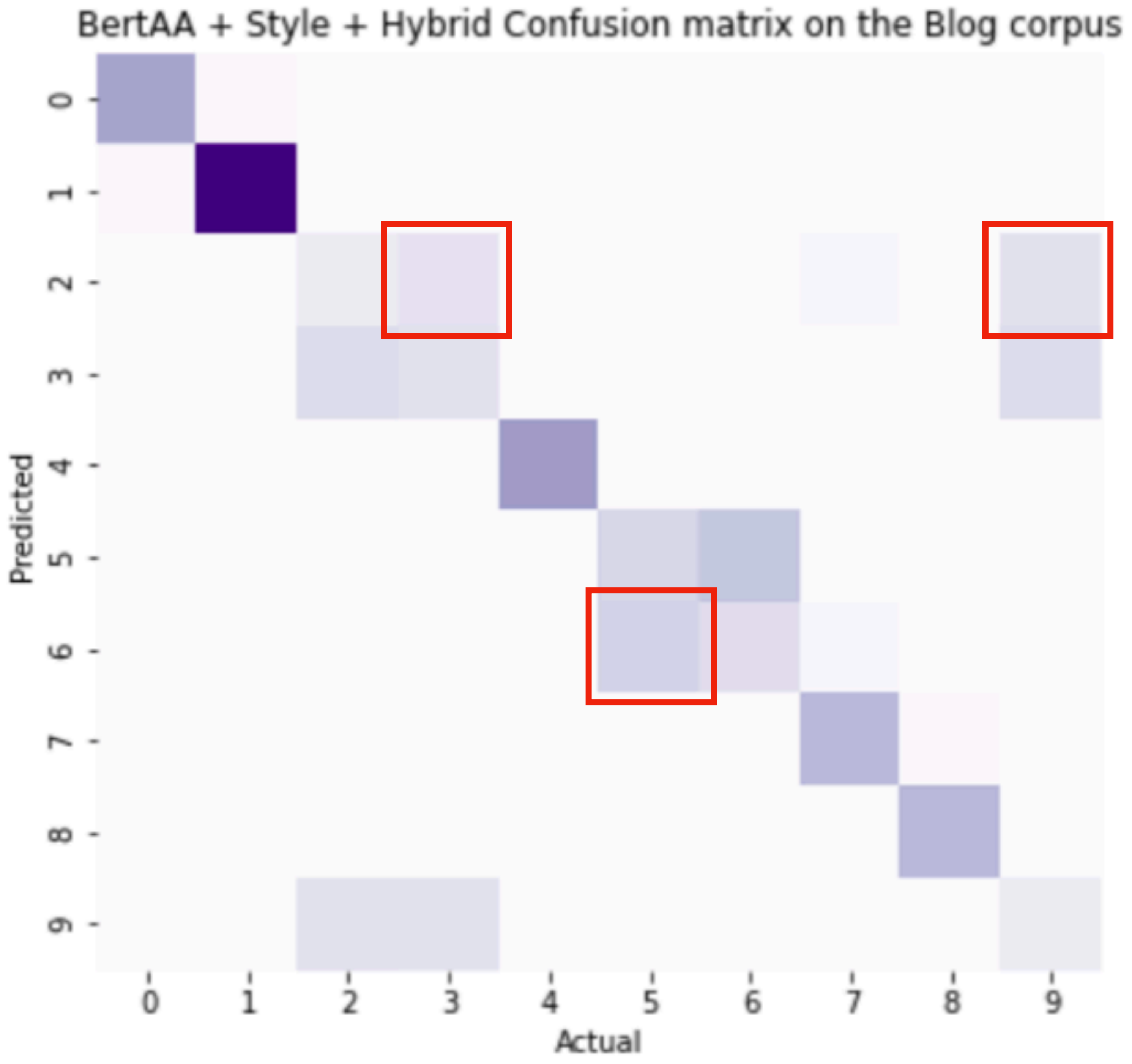
F1-Score improvement with external features

+**2.70**% with stylistic features

+**2.73**% with hybrid and stylistic features

Are external features useful?





Wider variety of errors

But errors are less important

What happens with less training data?

Approach	Accuracy
LDA+Hellinger (El et al.)	82
Word Level TF-IDF	91.4
CNN-Char (Ruder et al., 2016)	91.7
Comp.Att.+Sep.Rec. (Song et al., 2019)	91.8
Token-SVM (Seroussi et al., 2014)	92.52
SCAP (Frantzeskou et al., 2006)	94.8
Cont. N-gram Char (Sari et al., 2017)	94.8
(C+W+POS)/LM (Kamps et al., 2017)	95.9
N-gram + Style (Sari et al., 2018)	95.9
Syntax CNN(Zhang et al., 2018)	**96.2**
BertAA + Style + Hybrid - 1 epoch	88.7
BertAA + Style - 3 epochs	91.1
BertAA + Style + Hybrid - 5 epochs	92.3
BertAA + Style + Hybrid - 10 epochs	93.0

Accuracy on the IMDb62 Corpus

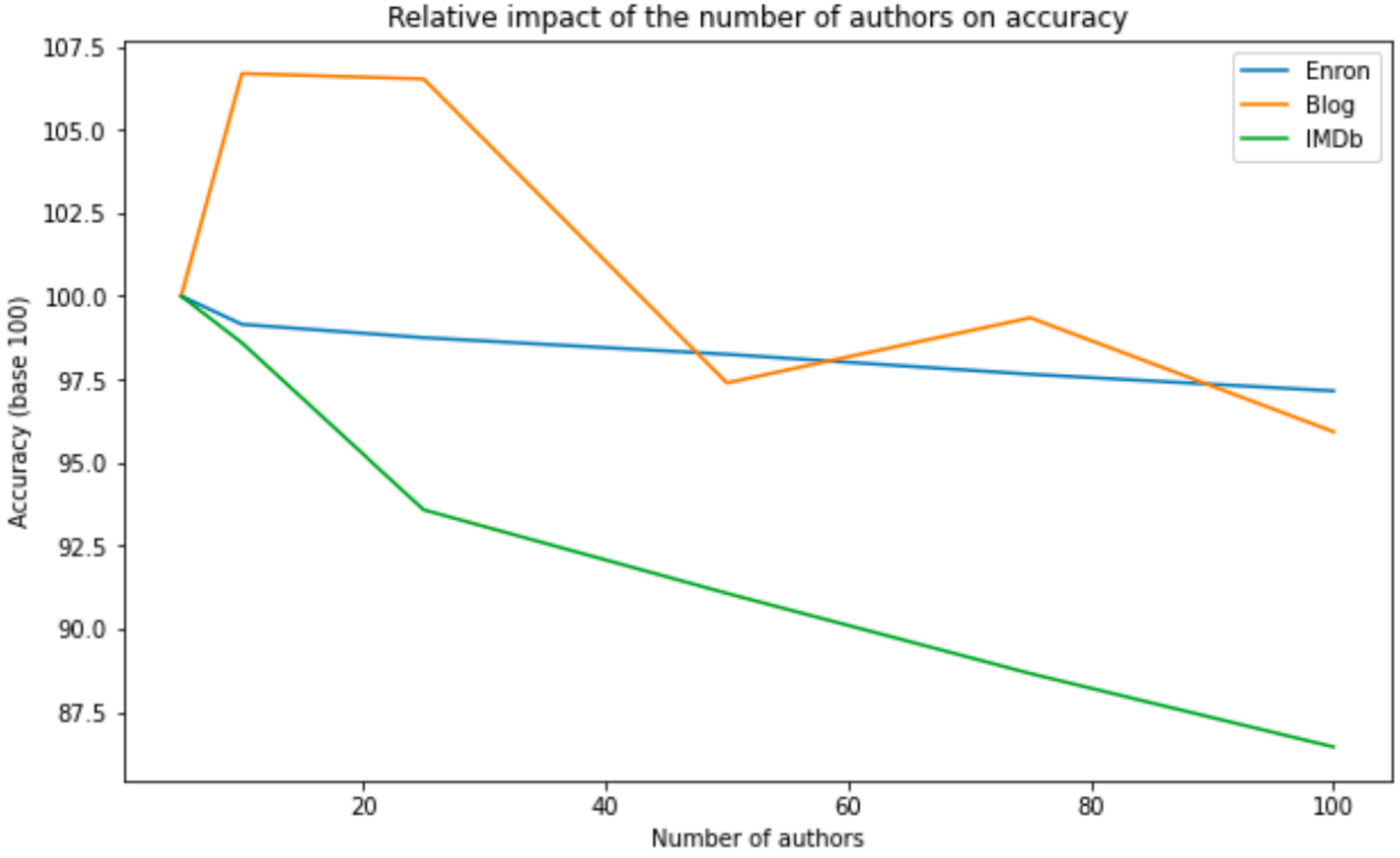
1000 texts per author

341 tokens on average

Longer and fewer texts

Performance below CNN

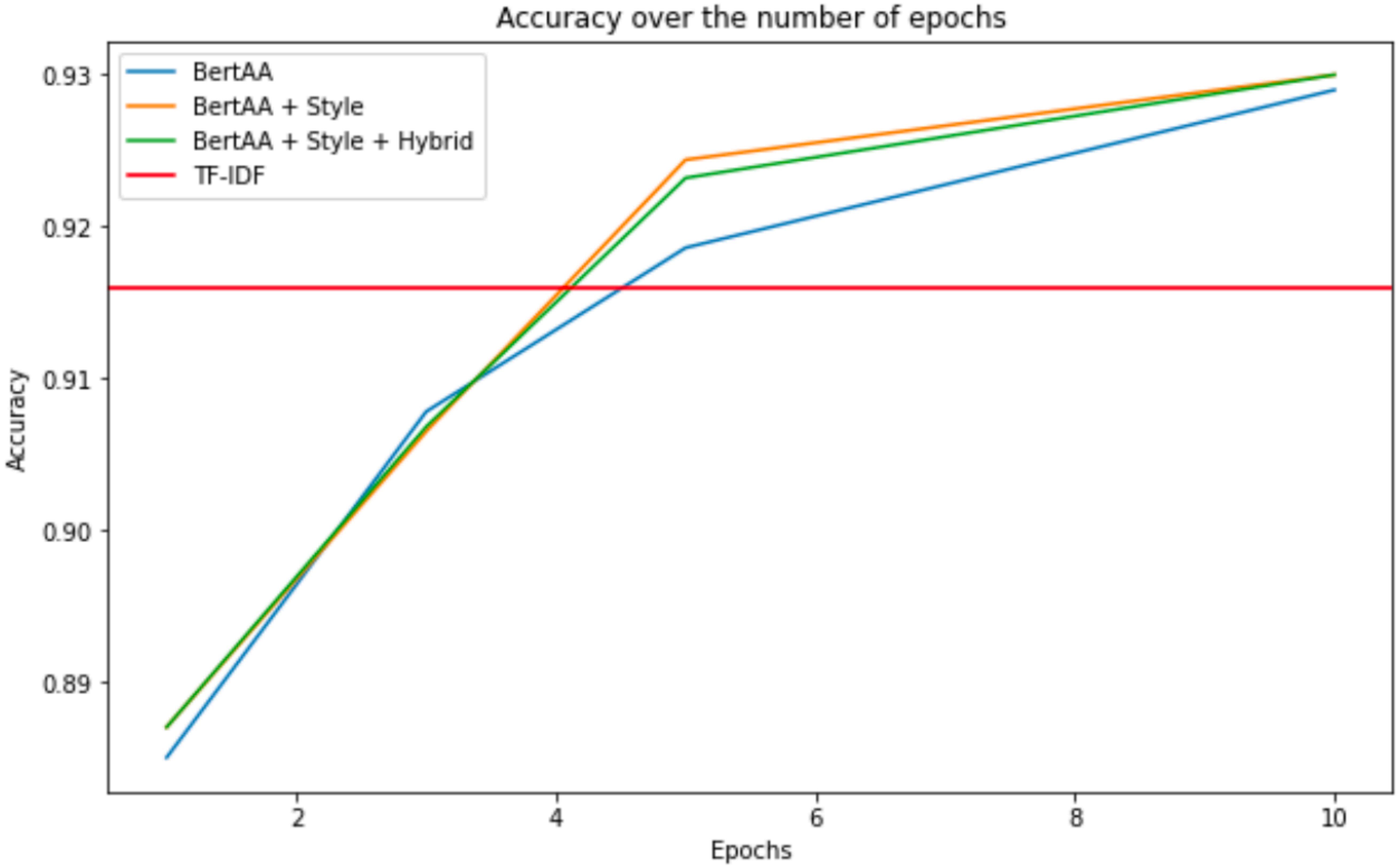
What happens with a more authors?



93% of the accuracy at 5 authors

maintained at 100 authors

How much fine-tuning is needed?



Accuracy on the IMDb62 Corpus

- Accuracy kept improving with the fine-tuning
- 5 epochs is a good trade-off with the time of fine-tuning

Take away message

Is the number of texts per author >> 1000?

Yes

No

How long are the texts?

Feature extraction + Classifier
CNNs

Twitter style

100 tokens

Long

Feature extraction + Classifier
Or CNNs

Metric

Syntax enriched CNN

F1-Score

Accuracy

BertAA + Style
(+Hybrid)

BertAA

Future works

- Further pre-training of BERT on target domain
- Explore other pre-trained Language Models
- Add new types of features
- Authorship Verification via similarity metrics on the embeddings
- Authorship Attribution on Automatic Speech Recognition transcripts in criminal investigations

Conclusion

- A BERT fine-tuning for AA
- That works well for a large number of texts
- And can be extended with external features to improve F1-score
- While setting a new SOTA on the Blog authorship dataset
- And a first benchmark on the full IMDb corpus

Datasets and code

Here

Contact

Maël Fabien

Ph.D. student at Idiap
Research Institute and EPFL

mael.fabien@idiap.ch

https://maelfabien.github.io/

https://github.com/maelfabien

https://www.linkedin.com/in/mael-fabien/

https://twitter.com/mael2ml

References

Traditional Methods:

- Lukas Muttenthaler, Gordon Lucas, and Janek Amann. « Authorship Attribution in Fan-Fictional Texts given variable length Character and Word N-Grams »
- Yunita Sari, Mark Stevenson, and Andreas Vlachos. 2018. « Topic or Style? Exploring the Most Useful Features for Authorship Attribution » in Proceedings of the 27th International Conference on Computational Linguistics, pages 343–353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- David Madigan, Alexander Genkin, David D. Lewis, and Dmitriy Fradkin. 2005. « Bayesian Multinomial Logistic Regression for Author Identification ». AIP Conference Proceedings, 803(1):509–516. Publisher: American Institute of Physics.
- Andrea Bacciu, Massimo La Morgia, Alessandro Mei, Eugenio Nerio Nemmi, and Julinda Stefa. 2020. « Cross-Domain Authorship Attribution Combining Instance-Based and Profile-Based Features ». page 14.

Deep Learning Methods:

- Chen Qian, Tianchang He, and Rao Zhang. « Deep Learning based Authorship Identification ». page 9
- Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2016. « Character-level and Multi-channel Convolutional Neural Networks for Large-scale Authorship Attribution ». arXiv:1609.06686 [cs]. ArXiv: 1609.06686.
- Prasha Shrestha, Sebastian Sierra, Fabio Gonzalez, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017. « Convolutional Neural Networks for Author- ship Attribution of Short Texts ». In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 669–674, Valencia, Spain. Association for Computational Linguistics.