Deep Bayes Factor Scoring for Authorship Verification

Benedikt Boenninghoff
Julian Rupp

Dorothea Kolossa
Robert M. Nickel*

PAN@CLEF2020

RUHR UNIVERSITÄT BOCHUM

*Bucknell UNIVERSITY

Authorship verification (AV) tasks at PAN 2020 to 2022[1] (Kestemont, Manjavacas, et al. 2020)

Task: Given two documents, determine if they were written by the same person

[1] https://pan.webis.de/clef20/pan20-web/author-identification.html

Authorship verification (AV) tasks at PAN 2020 to 2022[1] (Kestemont, Manjavacas, et al. 2020)

Task: Given two documents, determine if they were written by the same person

- **PAN 2020:** Closed-set / cross-fandom verification

[1] https://pan.webis.de/clef20/pan20-web/author-identification.html

Authorship verification (AV) tasks at PAN 2020 to 2022[1] (Kestemont, Manjavacas, et al. 2020)

Task: Given two documents, determine if they were written by the same person

- **PAN 2020:** Closed-set / cross-fandom verification
  - A large training dataset is provided by the PAN organizers (Bischoff, Deckers, et al. 2020)

[1] https://pan.webis.de/clef20/pan20-web/author-identification.html

Authorship verification (AV) tasks at PAN 2020 to 2022[1] (Kestemont, Manjavacas, et al. 2020)

Task: Given two documents, determine if they were written by the same person

- **PAN 2020:** Closed-set / cross-fandom verification
  - A large training dataset is provided by the PAN organizers (Bischoff, Deckers, et al. 2020)
  - Test set represents a subset of the authors/fandoms found in the training data

[1] https://pan.webis.de/clef20/pan20-web/author-identification.html

Authorship verification (AV) tasks at PAN 2020 to 2022[1] (Kestemont, Manjavacas, et al. 2020)

Task: Given two documents, determine if they were written by the same person

- **PAN 2020:** Closed-set / cross-fandom verification
  - A large training dataset is provided by the PAN organizers (Bischoff, Deckers, et al. 2020)
  - Test set represents a subset of the authors/fandoms found in the training data
- **PAN 2021:** Open-set verification

[1] https://pan.webis.de/clef20/pan20-web/author-identification.html

Authorship verification (AV) tasks at PAN 2020 to 2022[1] (Kestemont, Manjavacas, et al. 2020)

Task: Given two documents, determine if they were written by the same person

- **PAN 2020:** Closed-set / cross-fandom verification
  - A large training dataset is provided by the PAN organizers (Bischoff, Deckers, et al. 2020)
  - Test set represents a subset of the authors/fandoms found in the training data
- **PAN 2021:** Open-set verification
  - Test set now only contains "unseen" authors/fandoms

[1] https://pan.webis.de/clef20/pan20-web/author-identification.html

Authorship verification (AV) tasks at PAN 2020 to 2022[1] (Kestemont, Manjavacas, et al. 2020)

Task: Given two documents, determine if they were written by the same person

- **PAN 2020:** Closed-set / cross-fandom verification
  - A large training dataset is provided by the PAN organizers (Bischoff, Deckers, et al. 2020)
  - Test set represents a subset of the authors/fandoms found in the training data
- **PAN 2021:** Open-set verification
  - Test set now only contains "unseen" authors/fandoms
  - Training datset is identical to year one

[1] https://pan.webis.de/clef20/pan20-web/author-identification.html

Authorship verification (AV) tasks at PAN 2020 to 2022[1] (Kestemont, Manjavacas, et al. 2020)

Task: Given two documents, determine if they were written by the same person

- **PAN 2020:** Closed-set / cross-fandom verification
  - A large training dataset is provided by the PAN organizers (Bischoff, Deckers, et al. 2020)
  - Test set represents a subset of the authors/fandoms found in the training data
- **PAN 2021:** Open-set verification
  - Test set now only contains "unseen" authors/fandoms
  - Training datset is identical to year one
- **PAN 2022:** Role of judges at court

[1] https://pan.webis.de/clef20/pan20-web/author-identification.html

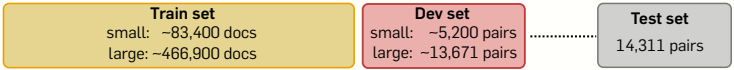
Text preprocessing strategies: **Preparing train/dev sets**

- Splitting the dataset into a train and a dev set[2]

[2] Dataset available at https://zenodo.org/record/3724096#.X2itQ3UzbQ8

Text preprocessing strategies: **Preparing train/dev sets**

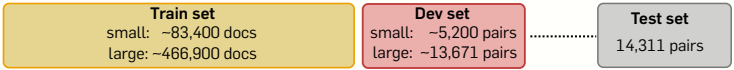
- Splitting the dataset into a train and a dev set[2]
- Removing all documents in the train set which also appear in the dev set



[2] Dataset available at https://zenodo.org/record/3724096#.X2itQ3UzbQ8

Text preprocessing strategies: **Topic Masking**

- Splitting the dataset into a train and a dev set[2]
- Removing all documents in the train set which also appear in the dev set
- Tokenizing (train/dev sets)[3] and counting words/characters (train set)

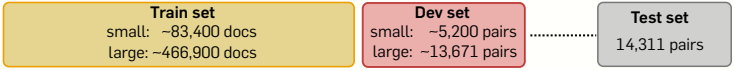


[2] Dataset available at https://zenodo.org/record/3724096#.X2itQ3UzbQ8

[3] Spacy tokenizer: https://spacy.io/

Text preprocessing strategies: **Topic Masking**

- Splitting the dataset into a train and a dev set[2]
- Removing all documents in the train set which also appear in the dev set
- Tokenizing (train/dev sets)[3] and counting words/characters (train set)
- Reducing the vocabulary sizes[4] : Mapping all rare token/character types to a special unknown symbol



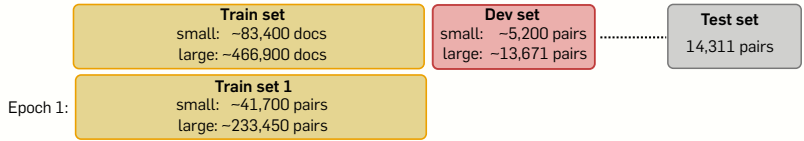
[2] Dataset available at `https://zenodo.org/record/3724096#.X2itQ3UzbQ8`

[3] Spacy tokenizer: `https://spacy.io/`

[4] Similar to text distortion algorithm 1 proposed in (Stamatatos 2017)

Text preprocessing strategies: **Data augmentation**

- Splitting the dataset into a train and a dev set[2]
- Removing all documents in the train set which also appear in the dev set
- Tokenizing (train/dev sets)[3] and counting words/characters (train set)
- Reducing the vocabulary sizes[4] : Mapping all rare token/character types to a special unknown symbol
- Re-sampling the pairs for train set in every epoch (Boenninghoff, Hessler, et al. 2019)



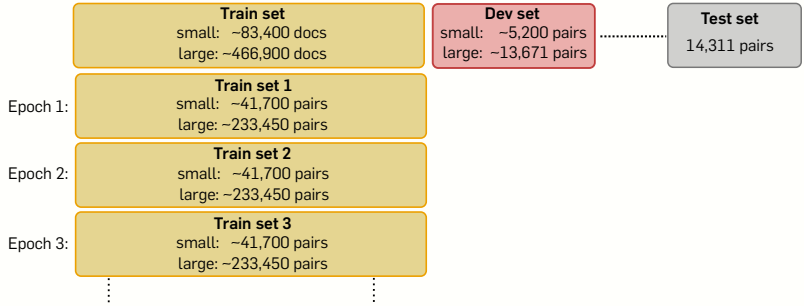
[2] Dataset available at `https://zenodo.org/record/3724096#.X2itQ3UzbQ8`

[3] Spacy tokenizer: `https://spacy.io/`

[4] Similar to text distortion algorithm 1 proposed in (Stamatatos 2017)

Text preprocessing strategies: **Data augmentation**

- Splitting the dataset into a train and a dev set[2]
- Removing all documents in the train set which also appear in the dev set
- Tokenizing (train/dev sets)[3] and counting words/characters (train set)
- Reducing the vocabulary sizes[4] : Mapping all rare token/character types to a special unknown symbol
- Re-sampling the pairs for train set in every epoch (Boenninghoff, Hessler, et al. 2019)
- Keeping all dev set pairs fixed!



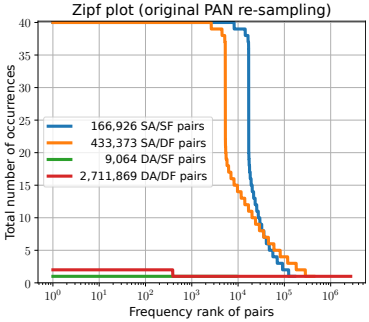
[2] Dataset available at `https://zenodo.org/record/3724096#.X2itQ3UzbQ8`

[3] Spacy tokenizer: `https://spacy.io/`

[4] Similar to text distortion algorithm 1 proposed in (Stamatatos 2017)

Improved re-sampling of document pairs[5]

- Problem: During training, our model repeatedly sees the same SA-pairs



[5] SA: same author, DA: different authors, SF: same fandom, DF: different fandoms

Improved re-sampling of document pairs[5]

- Modify the re-sampling of pairs w.r.t authorship and topical category

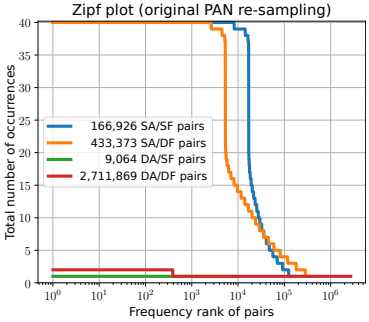
Algorithm 1 Re-sampling pairs

```
while authors with documents are available do
    for all authors do
        if r_1 ~ U[0, 1] < 1/2 then
            if r_2 ~ U[0, 1] < 1/2 then
                Try to sample SA/SF pair
            else
                Try to sample SA/DF pair
        else
            Try to sample a document for DA pairs
        Delete author from list if all documents are sampled
while two documents are available do
    if r_3 ~ U[0, 1] < 1/2 then
        Try to sample DA/SF pair
    else
        Try to sample DA/DF pair
```

SA vs. DA

SA/SF vs. SA/DF

DA/SF vs. DA/DF



[5] SA: same author, DA: different authors, SF: same fandom, DF: different fandoms

Improved re-sampling of document pairs[5]

- Modify the re-sampling of pairs w.r.t authorship and topical category

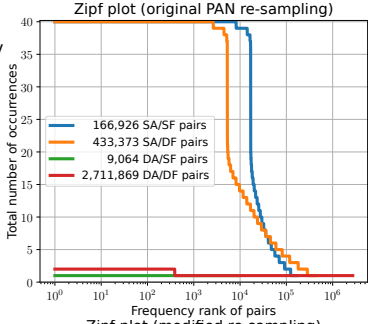
Algorithm 1 Re-sampling pairs

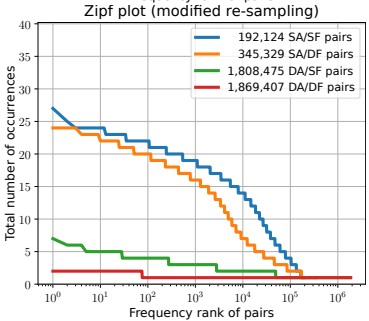
```
while authors with documents are available do
    for all authors do
        if r1 ~ U[0, 1] < 1/2 then
            if r2 ~ U[0, 1] < 1/2 then
                Try to sample SA/SF pair
            else
                Try to sample SA/DF pair
        else
            Try to sample a document for DA pairs
        Delete author from list if all documents are sampled
while two documents are available do
    if r3 ~ U[0, 1] < 1/2 then
        Try to sample DA/SF pair
    else
        Try to sample DA/DF pair
```

SA vs. DA

SA/SF vs. SA/DF

DA/SF vs. DA/DF





[5] SA: same author, DA: different authors, SF: same fandom, DF: different fandoms

Text preprocessing strategies: **(Overlapping) sliding windows with contextual prefixes**

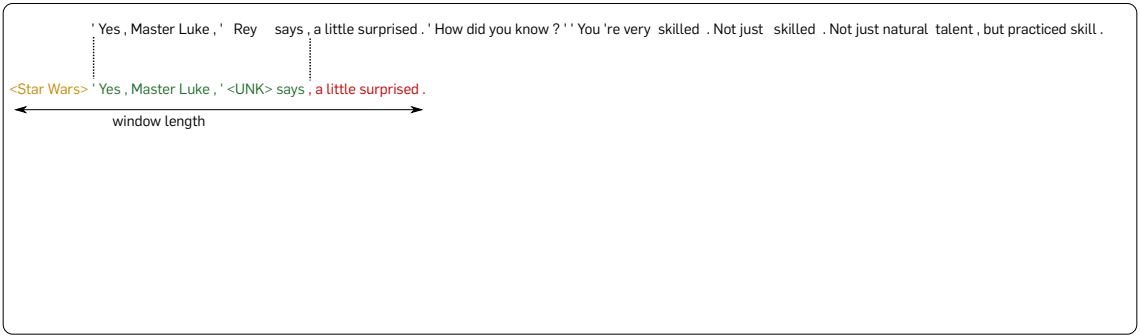
- Construct a sentence-like unit consisting of tokens that are grammatically linked

' Yes , Master Luke , ' Rey says , a little surprised . ' How did you know ? ' ' You 're very skilled . Not just skilled . Not just natural talent , but practiced skill .

<Star Wars> ' Yes , Master Luke , ' <UNK> says , a little surprised .

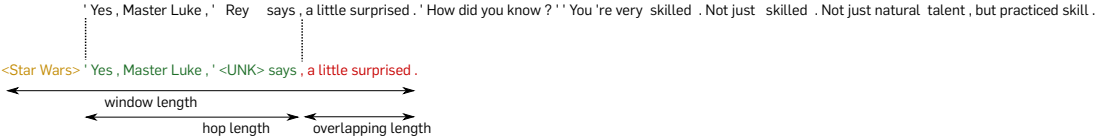
Text preprocessing strategies: **(Overlapping) sliding windows with contextual prefixes**

- Construct a sentence-like unit consisting of tokens that are grammatically linked



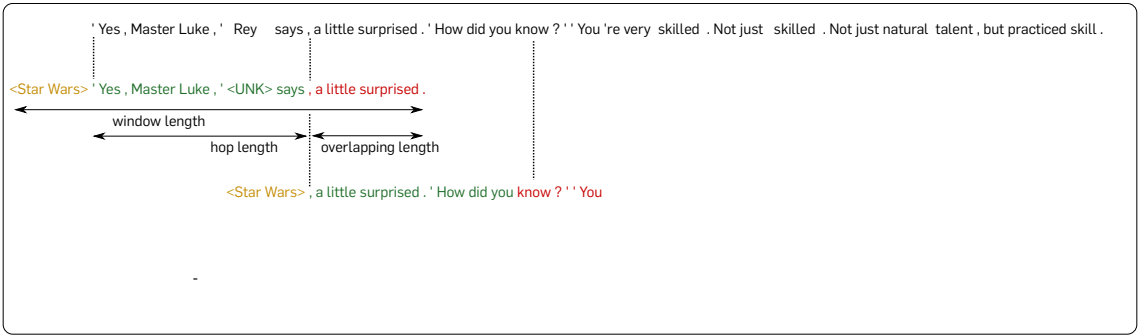
Text preprocessing strategies: **(Overlapping) sliding windows with contextual prefixes**

- Construct a sentence-like unit consisting of tokens that are grammatically linked



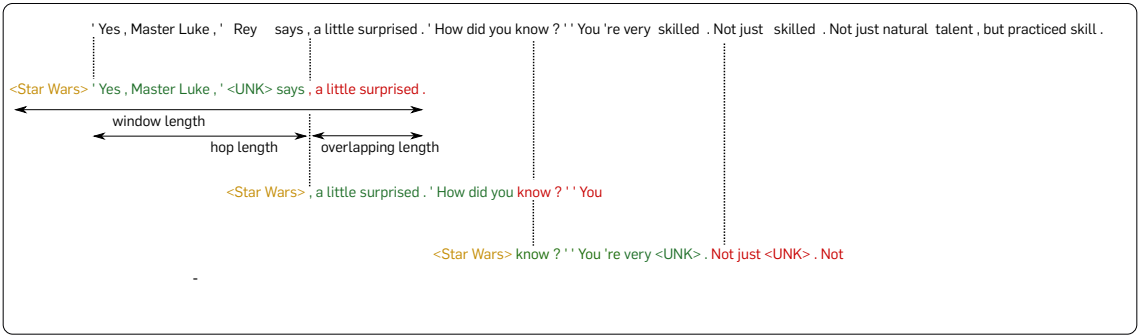
Text preprocessing strategies: **(Overlapping) sliding windows with contextual prefixes**

- Construct a sentence-like unit consisting of tokens that are grammatically linked
- `window_length = hop_length + overlapping_length + 1`



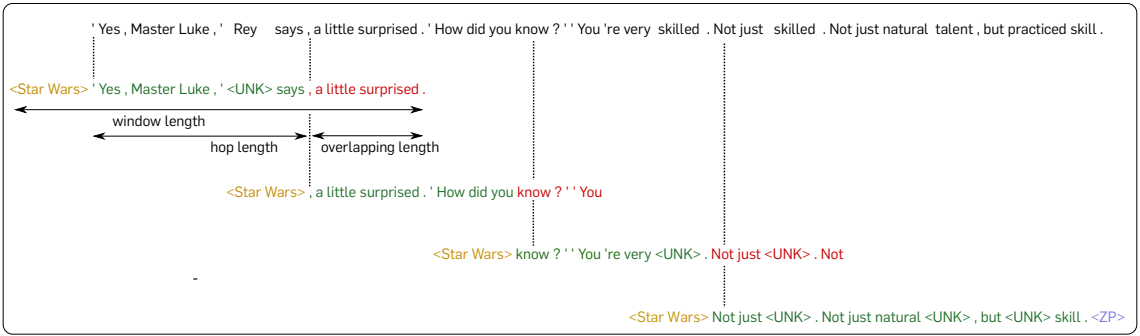
Text preprocessing strategies: **(Overlapping) sliding windows with contextual prefixes**

- Construct a sentence-like unit consisting of tokens that are grammatically linked
- `window_length = hop_length + overlapping_length + 1`

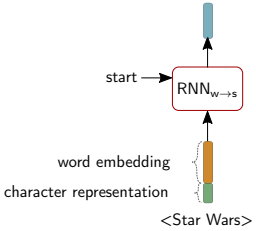


Text preprocessing strategies: **(Overlapping) sliding windows with contextual prefixes**

- Construct a sentence-like unit consisting of tokens that are grammatically linked
- `window_length = hop_length + overlapping_length + 1`

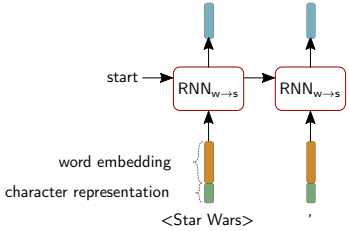


Hierarchical document encoding[6] (Boenninghoff, Nickel, et al. 2019)



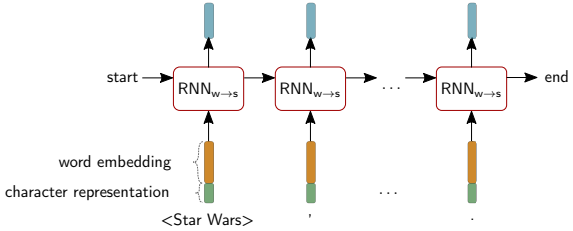
[6] Pretrained word embeddings taken from `https://fasttext.cc`

Hierarchical document encoding[6] (Boenninghoff, Nickel, et al. 2019)



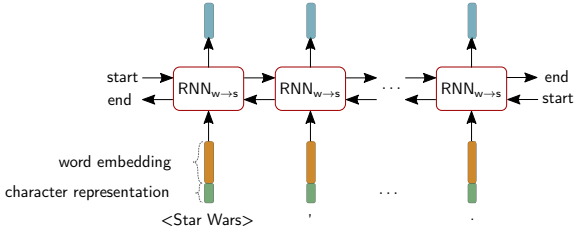
[6] Pretrained word embeddings taken from `https://fasttext.cc`

Hierarchical document encoding[6] (Boenninghoff, Nickel, et al. 2019)



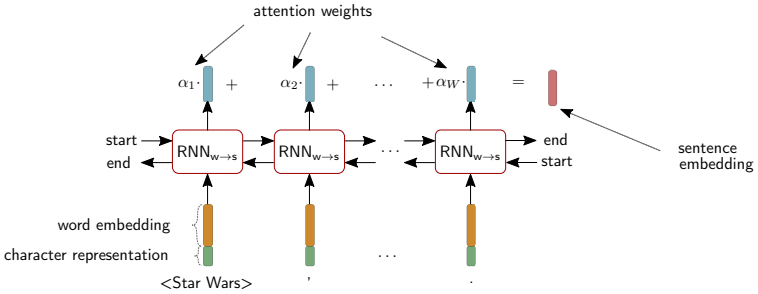
[6] Pretrained word embeddings taken from `https://fasttext.cc`

Hierarchical document encoding[6] (Boenninghoff, Nickel, et al. 2019)



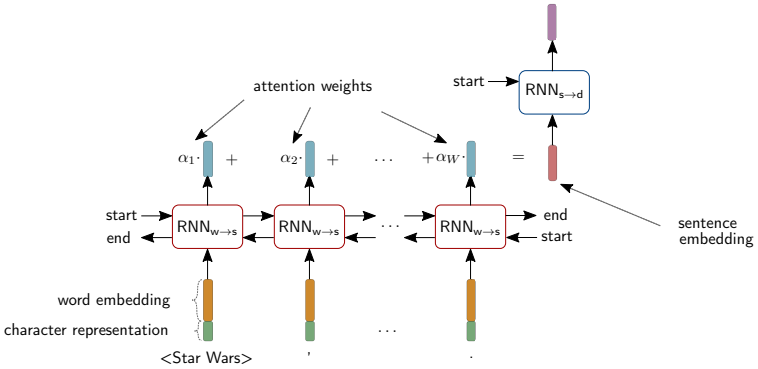
[6] Pretrained word embeddings taken from `https://fasttext.cc`

Hierarchical document encoding[6] (Boenninghoff, Nickel, et al. 2019)



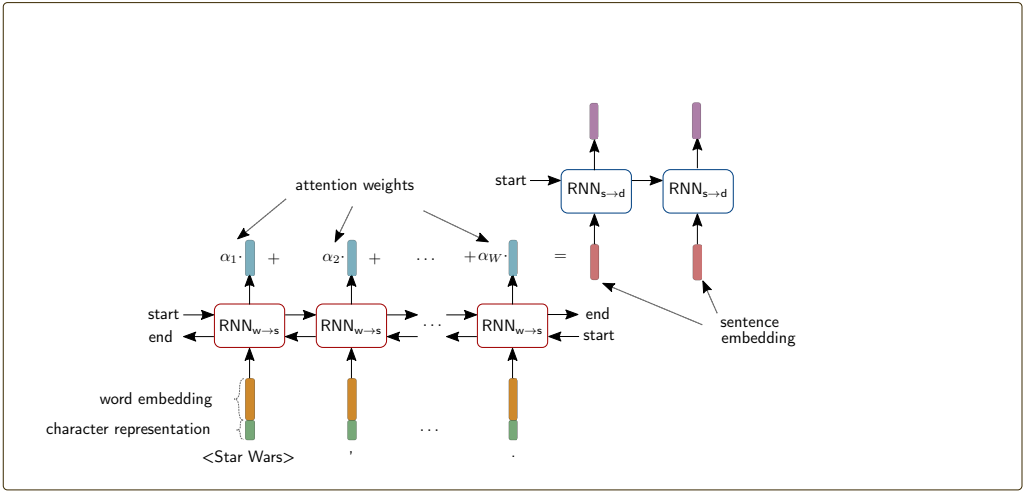
[6] Pretrained word embeddings taken from `https://fasttext.cc`

Hierarchical document encoding[6] (Boenninghoff, Nickel, et al. 2019)



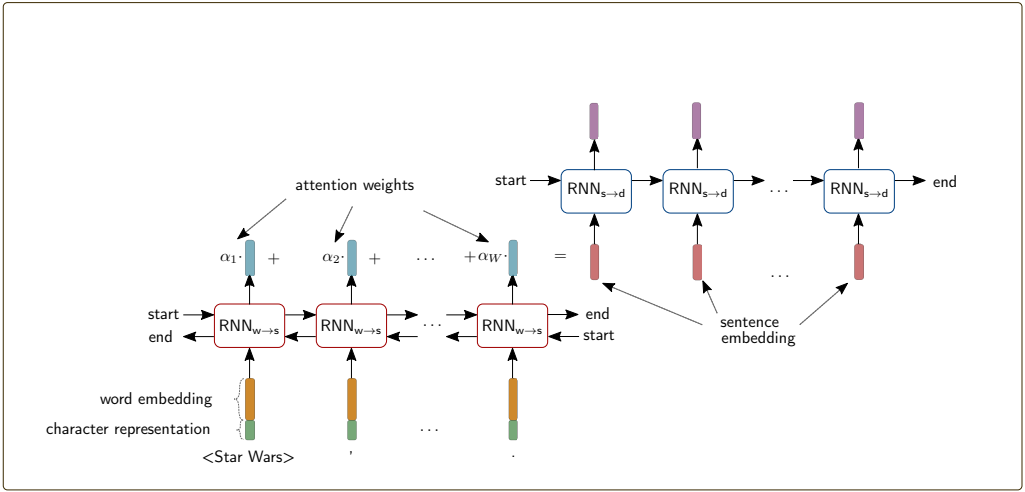
[6] Pretrained word embeddings taken from `https://fasttext.cc`

Hierarchical document encoding[6] (Boenninghoff, Nickel, et al. 2019)



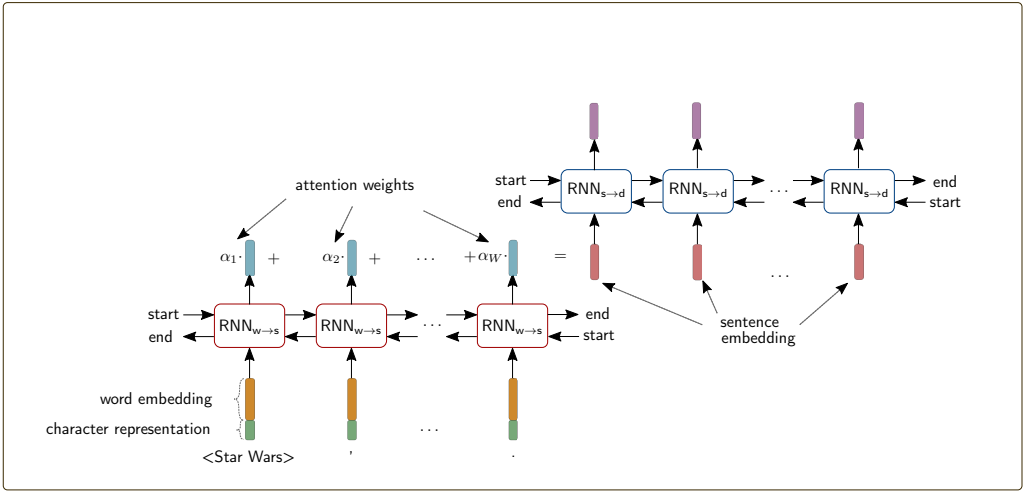
[6] Pretrained word embeddings taken from `https://fasttext.cc`

Hierarchical document encoding[6] (Boenninghoff, Nickel, et al. 2019)



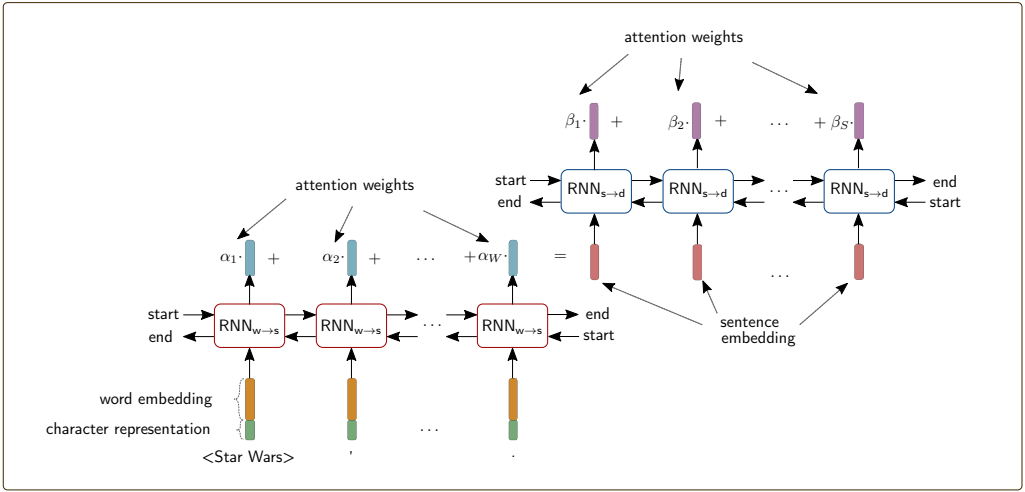
[6] Pretrained word embeddings taken from `https://fasttext.cc`

Hierarchical document encoding[6] (Boenninghoff, Nickel, et al. 2019)



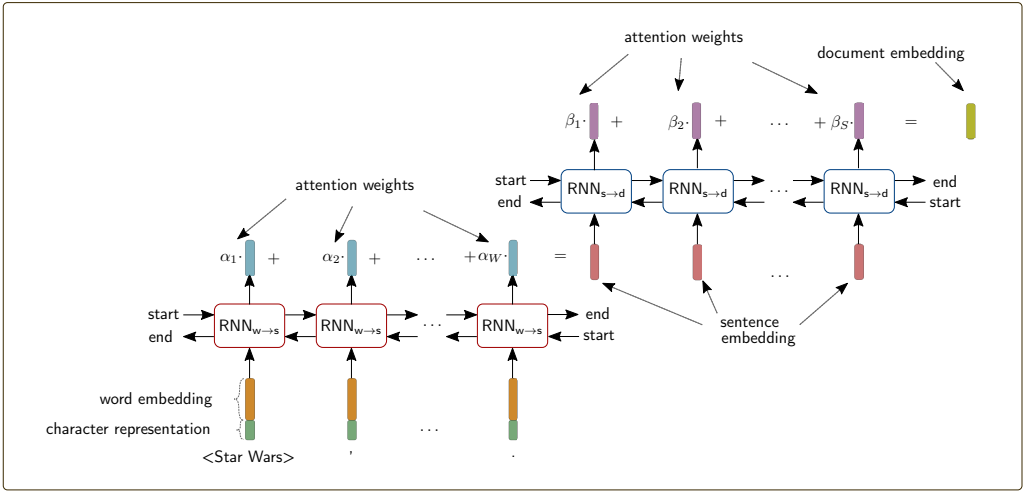
[6] Pretrained word embeddings taken from `https://fasttext.cc`

Hierarchical document encoding[6] (Boenninghoff, Nickel, et al. 2019)



[6] Pretrained word embeddings taken from `https://fasttext.cc`

Hierarchical document encoding[6] (Boenninghoff, Nickel, et al. 2019)



[6] Pretrained word embeddings taken from `https://fasttext.cc`

Deep Bayes factor scoring

- Define two hypotheses:

$\mathcal{H}_s$: Two documents were written by the same person

$\mathcal{H}_d$: Two documents were written by two different persons

Deep Bayes factor scoring

- Define two hypotheses:

$\mathcal{H}_s$: Two documents were written by the same person

$\mathcal{H}_d$: Two documents were written by two different persons

- Two-covariance model (Cumani, Brummer, et al. 2013):

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

with $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{B}^{-1})$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{W}^{-1})$

Deep Bayes factor scoring

- Define two hypotheses:

$\mathcal{H}_s$: Two documents were written by the same person

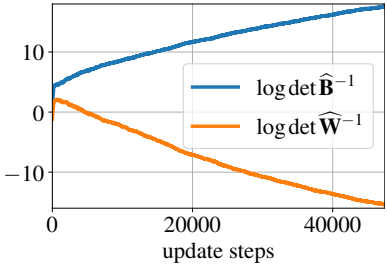
$\mathcal{H}_d$: Two documents were written by two different persons

- Two-covariance model (Cumani, Brummer, et al. 2013):

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

with $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{B}^{-1})$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{W}^{-1})$

Entropy curves during training:



Deep Bayes factor scoring

- Define two hypotheses:

$\mathcal{H}_s$: Two documents were written by the same person

$\mathcal{H}_d$: Two documents were written by two different persons

- Two-covariance model (Cumani, Brummer, et al. 2013):

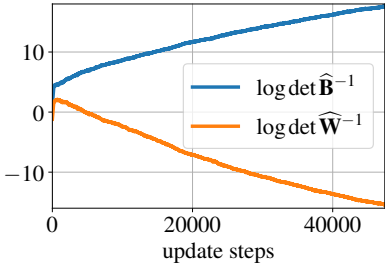
$$\underbrace{\boldsymbol{y}}_{\text{document embedding}} = \underbrace{\boldsymbol{x}}_{\text{author's writing style}} + \underbrace{\boldsymbol{\epsilon}}_{\text{noise term}}$$

with $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{B}^{-1})$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{W}^{-1})$

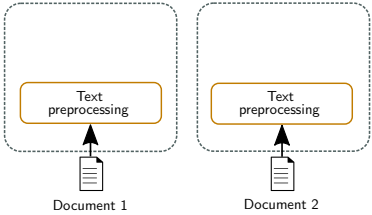
- Verification score:

$$\Pr(\mathcal{H}_s|\boldsymbol{y}_1, \boldsymbol{y}_2) = \frac{\Pr(\mathcal{H}_s)\, p(\boldsymbol{y}_1, \boldsymbol{y}_2|\mathcal{H}_s)}{\Pr(\mathcal{H}_s)\, p(\boldsymbol{y}_1, \boldsymbol{y}_2|\mathcal{H}_s) + \Pr(\mathcal{H}_d)\, p(\boldsymbol{y}_1, \boldsymbol{y}_2|\mathcal{H}_d)} \approx \frac{p(\boldsymbol{y}_1, \boldsymbol{y}_2|\mathcal{H}_s)}{p(\boldsymbol{y}_1, \boldsymbol{y}_2|\mathcal{H}_s) + p(\boldsymbol{y}_1, \boldsymbol{y}_2|\mathcal{H}_d)}$$

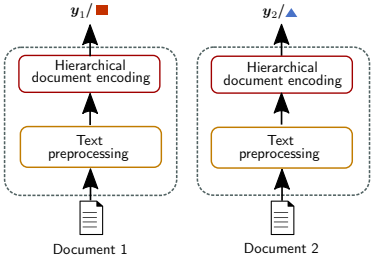
Entropy curves during training:



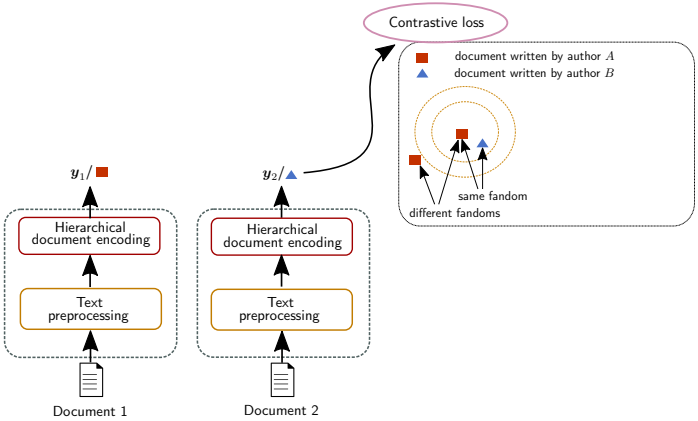
Combine binary cross-entropy and contrastive loss (Hu, Lu, and Tan 2014)



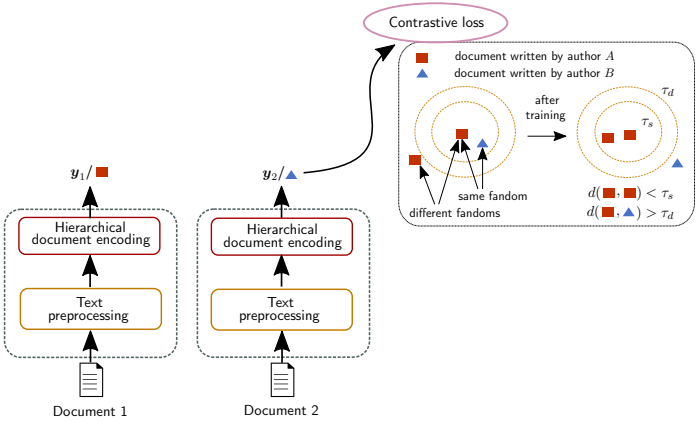
Combine binary cross-entropy and contrastive loss (Hu, Lu, and Tan 2014)



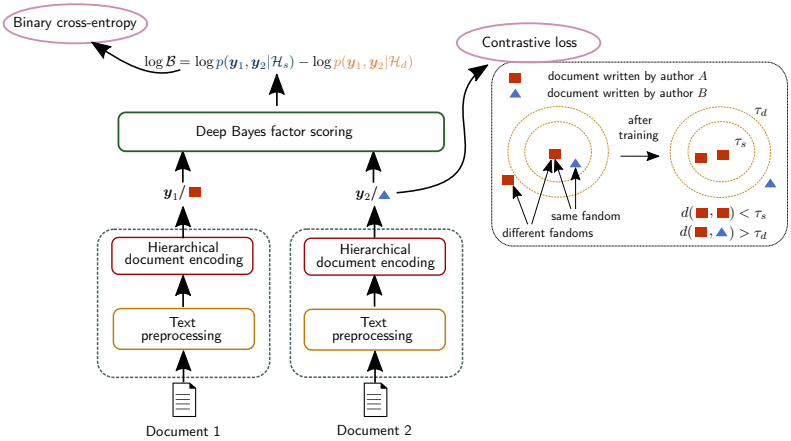
Combine binary cross-entropy and contrastive loss (Hu, Lu, and Tan 2014)



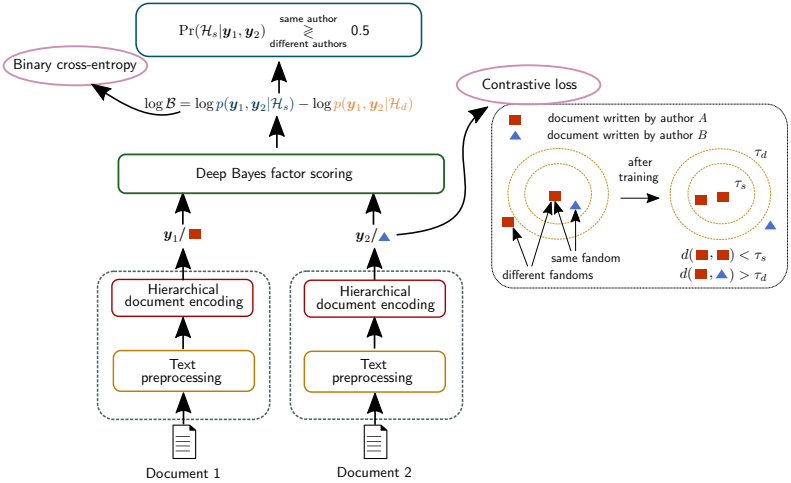
Combine binary cross-entropy and contrastive loss (Hu, Lu, and Tan 2014)



Combine binary cross-entropy and contrastive loss (Hu, Lu, and Tan 2014)



Combine binary cross-entropy and contrastive loss (Hu, Lu, and Tan 2014)



Evaluation results[7]

- Early-bird scores for dev set (small dataset)

		train set	evaluation	AUC	c@1	f_05_u	F1	overall
1	early-bird	small	dev set	0.964	0.919	0.916	0.932	0.933

[7] Colours represent the same models/runs

Evaluation results[7]

- Early-bird scores for test set ⇒ The model seems to generalize on the test set ☺

		train set	evaluation	AUC	c@1	f_05_u	F1	overall
1	early-bird	small	dev set	0.964	0.919	0.916	0.932	0.933
2	early-bird	small	test set	0.923	0.861	0.857	0.891	0.883

[7] Colours represent the same models/runs

Evaluation results[7]

- Best single runs for small/large datasets (at this step we introduced the contextual prefixes)

		train set	evaluation	AUC	c@1	f_05_u	F1	overall
1	early-bird	small	dev set	0.964	0.919	0.916	0.932	0.933
2	early-bird	small	test set	0.923	0.861	0.857	0.891	0.883
3	single	small	dev set	0.975	0.943	0.921	0.951	0.948
4	single	large	dev set	0.983	0.950	0.944	0.954	0.958

[7] Colours represent the same models/runs

Evaluation results[7]

- Ensembles that take the averaged vote from three independently trained "single" models

		train set	evaluation	AUC	c@1	f_05_u	F1	overall
1	early-bird	small	dev set	0.964	0.919	0.916	0.932	0.933
2	early-bird	small	test set	0.923	0.861	0.857	0.891	0.883
3	single	small	dev set	0.975	0.943	0.921	0.951	0.948
4	single	large	dev set	0.983	0.950	0.944	0.954	0.958
5	ensemble	small	dev set	0.977	0.942	0.938	0.946	0.951
6	ensemble	large	dev set	0.985	0.955	0.940	0.959	0.960

[7] Colours represent the same models/runs

Evaluation results[7]

- Results for ensembles on test set (including non-answers)

		train set	evaluation	AUC	c@1	f_05_u	F1	overall
1	early-bird	small	dev set	0.964	0.919	0.916	0.932	0.933
2	early-bird	small	test set	0.923	0.861	0.857	0.891	0.883
3	single	small	dev set	0.975	0.943	0.921	0.951	0.948
4	single	large	dev set	0.983	0.950	0.944	0.954	0.958
5	ensemble	small	dev set	0.977	0.942	0.938	0.946	0.951
6	ensemble	large	dev set	0.985	0.955	0.940	0.959	0.960
7	ensemble	small	test set	0.940	0.889	0.853	0.906	0.897
8	ensemble	large	test set	0.969	0.928	0.907	0.936	0.935

[7] Colours represent the same models/runs

Evaluation results[7]

- Model 9 = model 6/8 without defining non-answers

		train set	evaluation	AUC	c@1	f_05_u	F1	overall
1	early-bird	small	dev set	0.964	0.919	0.916	0.932	0.933
2	early-bird	small	test set	0.923	0.861	0.857	0.891	0.883
3	single	small	dev set	0.975	0.943	0.921	0.951	0.948
4	single	large	dev set	0.983	0.950	0.944	0.954	0.958
5	ensemble	small	dev set	0.977	0.942	0.938	0.946	0.951
6	ensemble	large	dev set	0.985	0.955	0.940	0.959	0.960
7	ensemble	small	test set	0.940	0.889	0.853	0.906	0.897
8	ensemble	large	test set	0.969	0.928	0.907	0.936	0.935
9	ensemble	large	test set	0.969	0.912	0.917	0.920	0.930

[7] Colours represent the same models/runs

Final ranking of the submitted approaches[8]

RANK	TEAM	TRAINING DATASET	AUC	C@1	F0.5U	F1-SCORE	OVERALL
1	boenninghoff20	large	0.969	0.928	0.907	0.936	0.935
2	weerasinghe20	large	0.953	0.880	0.882	0.891	0.902
3	boenninghoff20	small	0.940	0.889	0.853	0.906	0.897
4	weerasinghe20	small	0.939	0.833	0.817	0.860	0.862
5	halvani20b	small	0.878	0.796	0.819	0.807	0.825
6	kipnis20	small	0.866	0.801	0.815	0.809	0.823
7	araujo20	small	0.874	0.770	0.762	0.811	0.804
8	niven20	small	0.795	0.786	0.842	0.778	0.800
9	gagala20	small	0.786	0.786	0.809	0.800	0.796
10	araujo20	large	0.859	0.751	0.745	0.800	0.789
11	baseline (naive)	small	0.780	0.723	0.716	0.767	0.747
12	baseline (compression)	small	0.778	0.719	0.703	0.770	0.742
13	ordonez20	large	0.696	0.640	0.655	0.748	0.685
14	faber20	small	0.293	0.331	0.314	0.262	0.300

[8] https://pan.webis.de/clef20/pan20-web/author-identification.html

Looking forward to the PAN 2021 open-set AV challenge

- Simply splitting authors/fandoms into two disjoint groups

number of authors (train):	142,605
number of authors (dev):	29,543
number of fandoms (train):	1,120
number of fandoms (dev):	412

Looking forward to the PAN 2021 open-set AV challenge

- Simply splitting authors/fandoms into two disjoint groups
  - Train set: **136,068** pairs re-sampled in every epoch

number of authors (train):	**142,605**
number of authors (dev):	**29,543**
number of fandoms (train):	**1,120**
number of fandoms (dev):	**412**

Looking forward to the PAN 2021 open-set AV challenge

- Simply splitting authors/fandoms into two disjoint groups
  - Train set: **136,068** pairs re-sampled in every epoch
  - Dev set: **13,228** pairs

number of authors (train):	**142,605**
number of authors (dev):	**29,543**
number of fandoms (train):	**1,120**
number of fandoms (dev):	**412**

Looking forward to the PAN 2021 open-set AV challenge

- Simply splitting authors/fandoms into two disjoint groups
  - Train set: **136,068** pairs re-sampled in every epoch
  - Dev set: **13,228** pairs
- New challenging dev set:
  - It contains only "unseen" authors/fandoms
  - Cross-fandom orthogonality: Only SA/DF and DA/SF pairs

number of authors (train):	**142,605**
number of authors (dev):	**29,543**
number of fandoms (train):	**1,120**
number of fandoms (dev):	**412**

Looking forward to the PAN 2021 open-set AV challenge

- Simply splitting authors/fandoms into two disjoint groups
  - Train set: **136,068** pairs re-sampled in every epoch
  - Dev set: **13,228** pairs
- New challenging dev set:
  - It contains only "unseen" authors/fandoms
  - Cross-fandom orthogonality: Only SA/DF and DA/SF pairs
- First results (without non-answers and contextual prefixes):

number of authors (train):	142,605
number of authors (dev):	29,543
number of fandoms (train):	1,120
number of fandoms (dev):	412

	vocabulary size (characters)	vocabulary size (words)	hop_length	train word embeddings	AUC	c@1	f_05_u	F1	overall
1	150	15,000	25	YES	0.962	0.898	0.902	0.897	0.915
2	150	5,000	25	YES	0.969	0.907	0.909	0.906	**0.923**
3	150	50,000	25	YES	0.947	0.855	0.893	0.841	0.884
4	150	15,000	30	YES	0.961	0.896	0.903	0.894	0.913
5	750	15,000	25	YES	0.964	0.902	0.902	0.901	0.917
6	150	15,000	25	NO	0.962	0.896	0.905	0.894	0.914
7	150	5,000	25	NO	0.961	0.895	0.902	0.893	0.912

Conclusion and future work

Conclusion:

- AV models strongly depend on topical information (Kestemont, Manjavacas, et al. 2020)

Conclusion and future work

Conclusion:

- AV models strongly depend on topical information (Kestemont, Manjavacas, et al. 2020)
- Outstanding results achievable with traditional stylometric features (Weerasinghe and Greenstadt 2020)

Conclusion and future work

Conclusion:

- AV models strongly depend on topical information (Kestemont, Manjavacas, et al. 2020)
- Outstanding results achievable with traditional stylometric features (Weerasinghe and Greenstadt 2020)
- Surprisingly, BERT/Transformer-based models still do not outperform "traditional models" in this field
    - But very promising results in cross-domain authorship attribution (Barlas and Stamatatos 2020)

Conclusion and future work

Conclusion:

- AV models strongly depend on topical information (Kestemont, Manjavacas, et al. 2020)
- Outstanding results achievable with traditional stylometric features (Weerasinghe and Greenstadt 2020)
- Surprisingly, BERT/Transformer-based models still do not outperform "traditional models" in this field
  - But very promising results in cross-domain authorship attribution (Barlas and Stamatatos 2020)

Future work:

- Analysis of errors, contextual prefixes, re-sampling strategies, topic masking

Conclusion and future work

Conclusion:

- AV models strongly depend on topical information (Kestemont, Manjavacas, et al. 2020)
- Outstanding results achievable with traditional stylometric features (Weerasinghe and Greenstadt 2020)
- Surprisingly, BERT/Transformer-based models still do not outperform "traditional models" in this field
  - But very promising results in cross-domain authorship attribution (Barlas and Stamatatos 2020)

Future work:

- Analysis of errors, contextual prefixes, re-sampling strategies, topic masking
- Rethinking our handling of non-answers (e.g. Monte-Carlo dropout) on a calibration set

Conclusion and future work

Conclusion:

- AV models strongly depend on topical information (Kestemont, Manjavacas, et al. 2020)
- Outstanding results achievable with traditional stylometric features (Weerasinghe and Greenstadt 2020)
- Surprisingly, BERT/Transformer-based models still do not outperform "traditional models" in this field
  - But very promising results in cross-domain authorship attribution (Barlas and Stamatatos 2020)

Future work:

- Analysis of errors, contextual prefixes, re-sampling strategies, topic masking
- Rethinking our handling of non-answers (e.g. Monte-Carlo dropout) on a calibration set
- Transfer Learning: Incorporating contextualized word representations (e.g. ELMo, BERT)

Conclusion and future work

Conclusion:

- AV models strongly depend on topical information (Kestemont, Manjavacas, et al. 2020)
- Outstanding results achievable with traditional stylometric features (Weerasinghe and Greenstadt 2020)
- Surprisingly, BERT/Transformer-based models still do not outperform "traditional models" in this field
  - But very promising results in cross-domain authorship attribution (Barlas and Stamatatos 2020)

Future work:

- Analysis of errors, contextual prefixes, re-sampling strategies, topic masking
- Rethinking our handling of non-answers (e.g. Monte-Carlo dropout) on a calibration set
- Transfer Learning: Incorporating contextualized word representations (e.g. ELMo, BERT)
- Incorporating "compensation techniques" to deal with topical information
  - Domian-suppression (e.g. domain-adversarial training) (Bischoff, Deckers, et al. 2020)
  - Domian-adaptation (e.g. optimal transport) (Courty, Flamary, et al. 2017)

Conclusion and future work

Conclusion:

- AV models strongly depend on topical information (Kestemont, Manjavacas, et al. 2020)
- Outstanding results achievable with traditional stylometric features (Weerasinghe and Greenstadt 2020)
- Surprisingly, BERT/Transformer-based models still do not outperform "traditional models" in this field
  - But very promising results in cross-domain authorship attribution (Barlas and Stamatatos 2020)

Future work:

- Analysis of errors, contextual prefixes, re-sampling strategies, topic masking
- Rethinking our handling of non-answers (e.g. Monte-Carlo dropout) on a calibration set
- Transfer Learning: Incorporating contextualized word representations (e.g. ELMo, BERT)
- Incorporating "compensation techniques" to deal with topical information
  - Domian-suppression (e.g. domain-adversarial training) (Bischoff, Deckers, et al. 2020)
  - Domian-adaptation (e.g. optimal transport) (Courty, Flamary, et al. 2017)

Acknowledgement

Big thanks to the PAN2020-AV-team for organizing the shared task! ☺

References I

- Georgios Barlas and Efstathios Stamatatos. “Cross-Domain Authorship Attribution Using Pre-trained Language Models”. In: *Artificial Intelligence Applications and Innovations*. Ed. by Ilias Maglogiannis, Lazaros Iliadis, and Elias Pimenidis. Springer International Publishing, 2020, pp. 255–266.
- Sebastian Bischoff, Niklas Deckers, Marcel Schliebs, Ben Thies, Matthias Hagen, Efstathios Stamatatos, Benno Stein, and Martin Potthast. “The Importance of Suppressing Domain Style in Authorship Analysis”. In: *CoRR* abs/2005.14714 (2020).
- Benedikt Boenninghoff, Steffen Hessler, Dorothea Kolossa, and Robert M. Nickel. “Explainable Authorship Verification in Social Media via Attention-based Similarity Learning”. In: *2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, December 9-12, 2019*. IEEE, 2019, pp. 36–45.
- Benedikt Boenninghoff, Robert M. Nickel, Steffen Zeiler, and Dorothea Kolossa. “Similarity Learning for Authorship Verification in Social Media”. In: *Proc. ICASSP*. 2019, pp. 2457–2461.
- N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. “Optimal Transport for Domain Adaptation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.9 (2017), pp. 1853–1865.

References II

- Sandro Cumani, Niko Brummer, Lukáš Burget, Pietro Laface, Oldřich Plchot, and Vasileios Vasilakakis. “Pairwise Discriminative Speaker Verification in the I -Vector Space”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 2013.6 (2013), pp. 1217–1227.
- J. Hu, J. Lu, and Y. P. Tan. “Discriminative Deep Metric Learning for Face Verification in the Wild”. In: *Proc. CVPR*. 2014, pp. 1875–1882.
- Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. “Overview of the Cross-Domain Authorship Verification Task at PAN 2020”. In: *CLEF 2020 Labs and Workshops, Notebook Papers*. Ed. by Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol. CEUR-WS.org, 2020.
- Efstathios Stamatatos. “Authorship Attribution Using Text Distortion”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, 2017, pp. 1138–1149.
- Janith Weerasinghe and Rachel Greenstadt. “Feature Vector Difference based Neural Network and Logistic Regression Models for Authorship Verification—Notebook for PAN at CLEF 2020”. In: *CLEF 2020 Labs and Workshops, Notebook Papers*. Ed. by Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol. CEUR-WS.org, 2020.