ATHENE
National Research Center for Applied Cybersecurity

TAVeer - An Interpretable Topic-Agnostic Authorship Verification Method

Oren Halvani, Lukas Graner and Roey Regev (Fraunhofer SIT)

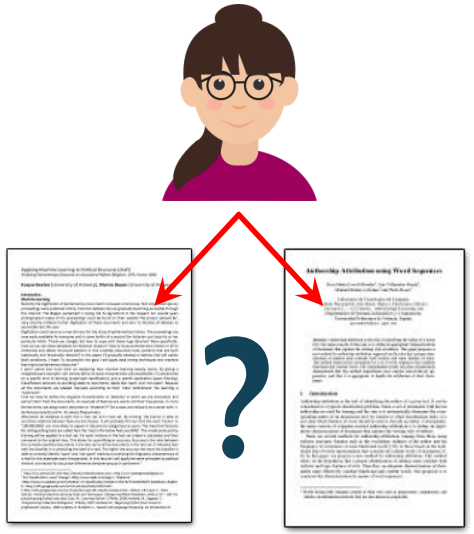
PAN: Stylometry and Digital Text Forensics held in conjunction with the CLEF 2020 Conference and Labs of the Evaluation Forum Information Access Evaluation meets Multilinguality, Multimodality, and Visualization, 22 - 25 September 2020, Thessaloniki - Greece

ATHENE is a research center of the Fraunhofer-Gesellschaft with the participation of

Fraunhofer SIT
Fraunhofer IGD
TECHNISCHE UNIVERSITÄT DARMSTADT
h_da HOCHSCHULE DARMSTADT UNIVERSITY OF APPLIED SCIENCES

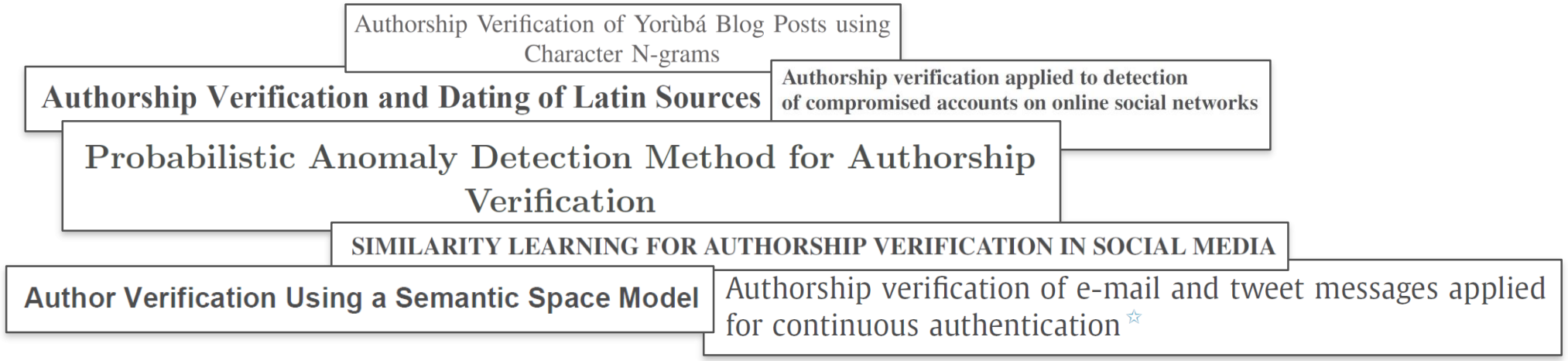
Motivation

- A central question that has occupied digital text forensics for decades is how to determine whether two documents were written by the same person
- Authorship verification (AV) is a branch of research that deals with this important question
- AV can be used for a wide range of applications including:
  - Continuous authentication
  - Expose malicious emails
  - Ghostwriting / plagiarism detection
  - Authentication of historical writings
  - Detection of speech changes in dementia patients
  - …



Motivation

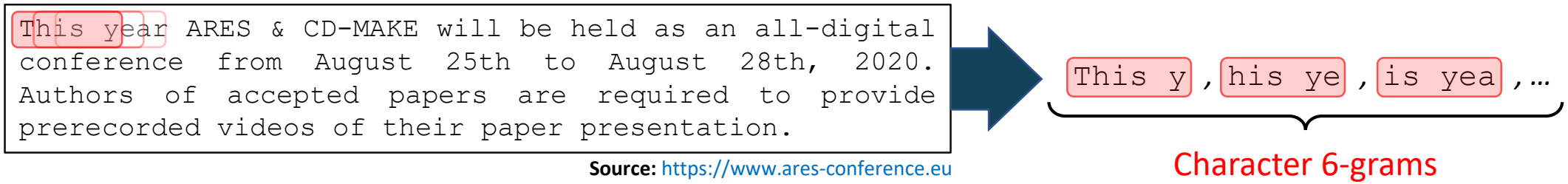
- Over the years, research activities in the field of AV have steadily increased, which has led to numerous approaches that aim to solve this problem e.g.



- However, a large number of existing AV approaches consider features within the documents that are not always related to the writing style...

Motivation

- Many AV methods, for example, rely on implicitly defined features such as character n-grams:



Source: https://www.ares-conference.eu

- Characters n grams are extracted uncontrolled from texts and thus capture text units that are not only related to the writing style but also to other document properties such as topic, genre, structure, sentiment, etc.

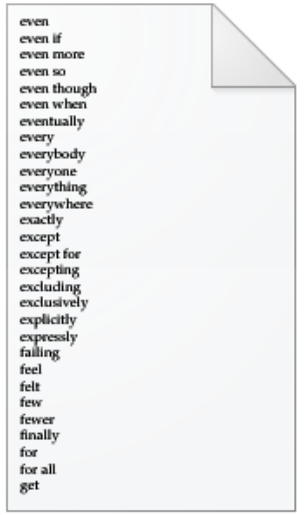
- Therefore, it may accidentally happen that the prediction of an AV method is not really based on the writing style, so that it will miss its true purpose

Proposed Feature Categories

- To counteract this, we follow an alternative idea in which we consider explicitly defined features. More precisely, we focus on 20 categories of topic-agnostic (**TA**) words and phrases…

Category	Examples
Conjunctions	{and, as, because, but, either, for, hence, however, if, neither, nor, once, …}
Determiners	{a, an, both, each, either, every, no, other, our, some, …}
Prepositions	{above, across, after, among, below, beside, between, beyond, inside, outside, …}
Pronouns	{all, another, any, anyone, anything, everything, few, he, her, hers, herself, …}
Quantifiers	{any, certain, each, either, few, less, lots, many, more, most, much, neither, …}
Auxiliary verbs	{can, could, might, must, ought, shall, will, …}
Delexicalised verbs	{get, go, take, make, do, have, give, set, …}
Empty verbs	{do, did, does, got, getting, have, had, had, gives, giving, gave, give, gets, …}
Helping verbs	{am, is, are, was, were, be, been, being, will, should, would, could, …}
Contractions	{i'm, i'd, i'll, i've, he's, it's, we'd, she's, it'll, we're, how's, you're, …}
Adverbs of degree	{almost, enough, hardly, just, nearly, quite, simply, so, too, …}
Adverbs of frequency	{again, always, never, normally, rarely, seldom, sometimes, usually, …}
Adverbs of place	{above, below, everywhere, here, in, inside, into, nowhere, out, outside, there, …}
Adverbs of time	{already, during, immediately, just, late, recently, still, then, sometimes, yet, …}
Pronominal adverbs	{hereafter, hereby, thereafter, thereby, therefore, therein, whereas, wherever, …}
Focusing adverbs	{especially, mainly, particularly, generally, only, simply, exactly, merely, solely, …}
Conjunctive adverbs	{likewise, meanwhile, moreover, namely, nonetheless, otherwise, perhaps, rather, …}
Transition words	{besides, furthermore, generally, hence, thus, however, incidentally, subsequently, …}
Transitional phrases	{of course, as a result, in addition, because of, in contrast, on the other hand, …}
Phrasal prepositions	{as opposed to, in regard to, in relation to, inspite of, out of, with regard to, …}

$\mathcal{L}_{TA}$



≈ 1000 words and phrases

Proposed Feature Categories

- Based on $\mathcal{L}_{TA}$, we propose the following feature categories that are used by our AV method

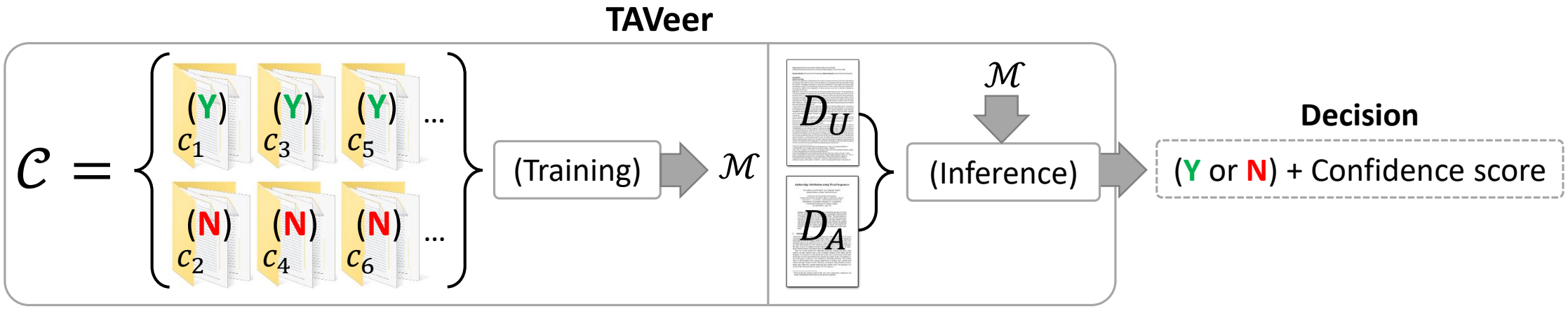
Example sentence: "So that's the way it goes."

ID	Feature category	Range	Sample n	Sample output
F_{1-3}	Punctuation n-grams	$n \in \{1, 2, 3\}$	$n = 2$	{('.)}
F_4	TA sentence and clause starters	—		{(so)}
F_5	TA sentence endings	—		{(goes)}
F_{6-9}	TA token n-grams	$n \in \{1, 2, 3, 4\}$	$n = 3$	{(so that's the), (it goes .)}
F_{10-11}	TA masked token n-grams	$n \in \{3, 4\}$	$n = 3$	{(that's the #), (the # it), (# it goes)}

- Note that here, unlike standard n-grams, we have full control over which text units are captured

Proposed AV Approach

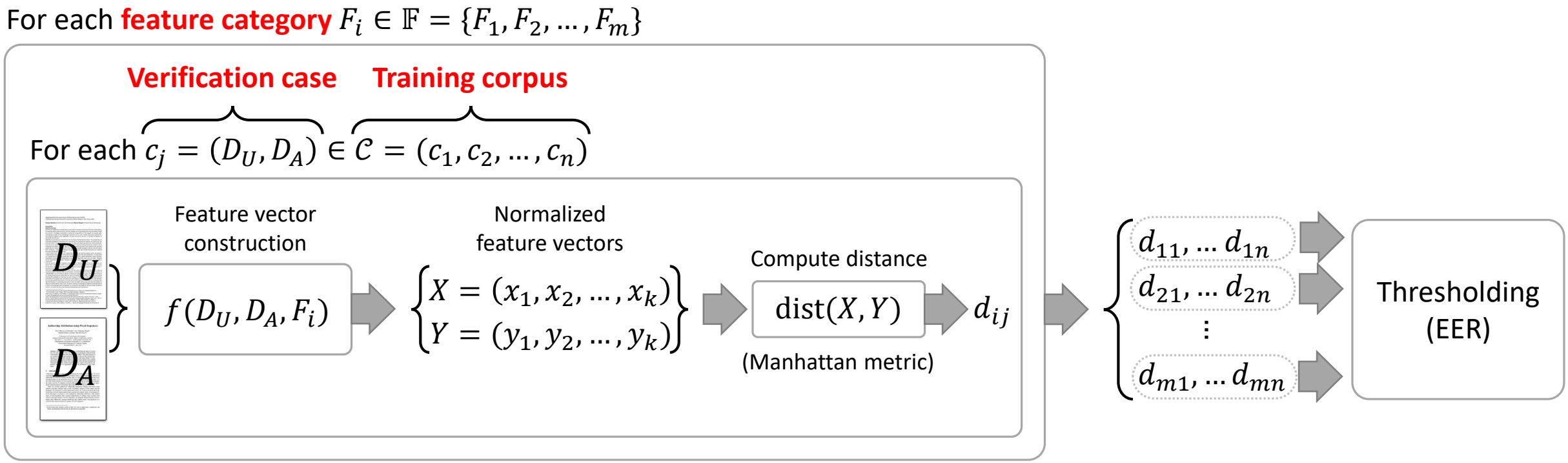
- Based on the proposed feature categories, we introduce in the following our alternative AV approach **TAVeer**
- TAVeer can essentially be divided into two phases: training and inference



- **Training**: A model $\mathcal{M}$ has to be "learned" on the basis of a training corpus $\mathcal{C}$ consisting of known verification cases labeled as **Y** (same author) and **N** (different author)
- **Inference**: $\mathcal{M}$ is applied to an unseen verification case in order to accept or reject the questioned authorship

TAVeer (Training)

- Required building blocks for the calculation of distances and thresholds…

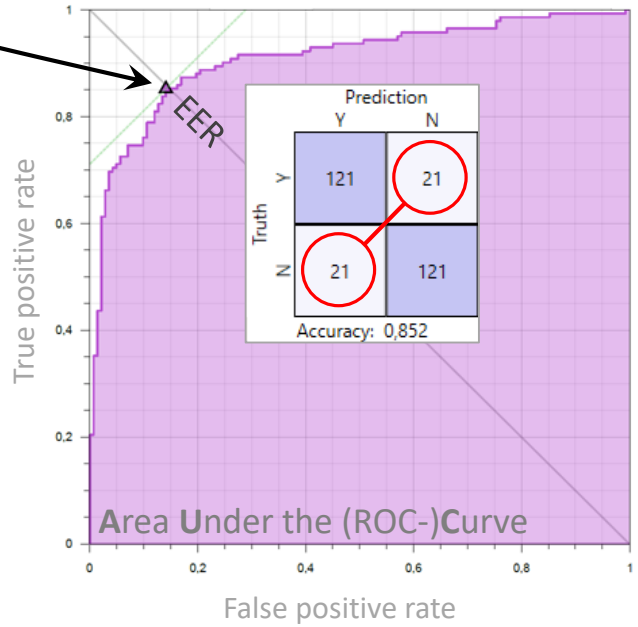


TAVeer (Training)

- After all distances have been computed, we determine for each F_i its corresponding threshold θ_{F_i} via the EER (**e**qual **e**rror **r**ate), which is the **point** on the ROC-curve where the false positive rate equals the false negative rate

- In our setting, all corpora are balanced (same number of **Y**/**N**-cases). Therefore, we use the median as an approximation of the EER

- The result of the thresholding procedure is a set $\Theta = \{\theta_{F_1}, \theta_{F_2}, \ldots, \theta_{F_m}\}$, with $\theta_{F_i} = \text{median}(d_{i1}, d_{i2}, \ldots, d_{in})$



TAVeer (Training)

- To construct $\mathcal{M}$, we further require a **similarity function** that:

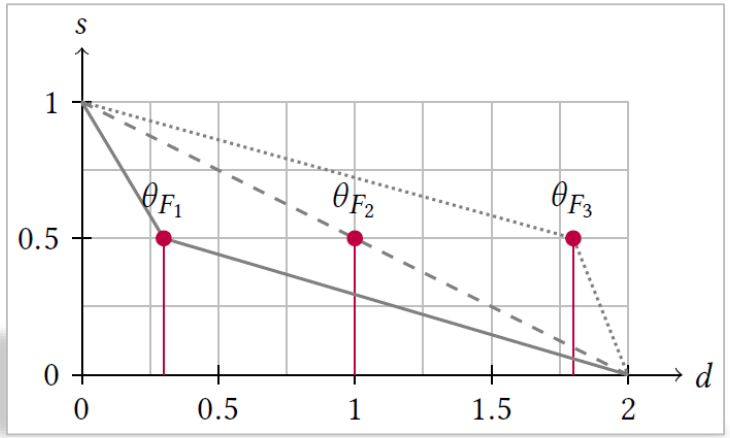
  (1) transforms the computed distances into similarity scores falling into $[0; 1]$ and

  (2) calibrates these scores so that 0.5 marks the decision boundary

- For the intended purpose, we designed the following piecewise function:

$$\mathrm{sim}(d, d_{max}, \theta_F) = \begin{cases} 1 - \dfrac{d}{2\theta_F}, & \text{if } d < \theta_F \\ \dfrac{1}{2} - \dfrac{d - \theta_F}{2(d_{max} - \theta_F)}, & \text{otherwise} \end{cases}$$

Upper bound of the distance function
(for the Manhattan metric $d_{max} = 2$ holds)



TAVeer (Training)

- To find a suitable ensemble, we first create a set $\mathbb{F}_\Theta = \{(F_1, \theta_{F_1}), (F_2, \theta_{F_2}), \ldots, (F_m, \theta_{F_m})\}$

- Based on $\mathbb{F}_\Theta$ we generate all possible ensembles $\mathcal{E}_1, \mathcal{E}_2$,… by using the powerset:

$$\mathcal{P}(\mathbb{F}_\Theta) \setminus \emptyset = \{\underbrace{\{(F_1, \theta_{F_1})\}}_{\text{Atomic ensemble}}, \underbrace{\{(F_1, \theta_{F_1}), (F_2, \theta_{F_2})\}}_{\text{Ensemble}}, \ldots\}$$

- Next, we construct an aggregated similarity function on top of $\mathrm{sim}(\cdot)$ to take $\mathcal{E}$ into account:

$$\mathrm{sim}_\mathcal{E}(D_U, D_A, d_{max}, \mathcal{E}) = \mathrm{median}(\{\mathrm{sim}(\underbrace{\mathrm{dist}(f(D_U, D_A, F))}_{\mathrm{dist}(X,Y)}, d_{max}, \theta_F) | (F, \theta_F) \in \mathcal{E}\})$$

TAVeer (Training)

- To find an optimal $\mathcal{E}$ that will be chosen as the final model $\mathcal{M}$, a ranking mechanism is needed…

- For this, we define a classification function:

$$\text{classify}(D_U, D_A, d_{max}, \mathcal{E}) = \begin{cases} \textbf{Y}\ (\text{same author}), & \text{if } \text{sim}_{\mathcal{E}}(D_U, D_A, d_{max}, \mathcal{E}) > 0.5 \\ \textbf{N}\ (\text{different author}), & \text{otherwise} \end{cases}$$

- Using this function, we classify all verification cases $(c_1, c_2, \ldots, c_n)$ in the training corpus $\mathcal{C}$ for each ensemble $\mathcal{E}_i \in \mathcal{P}(\mathbb{F}_\Theta)$ and calculate the respective accuracies

TAVeer (Training)

- To obtain the optimal $\mathcal{E}$, we sort all resulting ensembles one by one according to the following three criteria (each in descending order):

  (1) Accuracy of $\mathcal{E}$ (calculated for $\mathcal{C}$)

  (2) Number of feature categories $\mathcal{E}$ contains

  (3) Median accuracy regarding all atomic ensembles in $\mathcal{E}$ (calculated for $\mathcal{C}$)

- Finally, we select the first ensemble from the sorted list, which represents the final model $\mathcal{M}$

TAVeer (Inference)

- Based on the resulting model $\mathcal{M}$, TAVeer performs the following steps, to decide for an unseen verification case $c_{\text{new}} = (D_U, D_A)$ whether both documents were written by the same author

- Using $\text{classify}(\cdot)$, TAVeer first computes the aggregated similarity value:

$$s_{\text{new}} = \text{sim}_{\mathcal{C}}(D_U, D_A, d_{max}, \mathcal{M})$$

- Afterwards, a binary prediction (Y/N) regarding the questioned authorship of D_U is obtained by comparing s_{new} against the decision boundary...

$$\text{decision}(c_{\text{new}}) = \begin{cases} (\text{Y}, s_{\text{new}}), \text{if } s_{\text{new}} > 0.5 \\ (\text{N}, s_{\text{new}}), \text{otherwise} \end{cases}$$

Evaluation

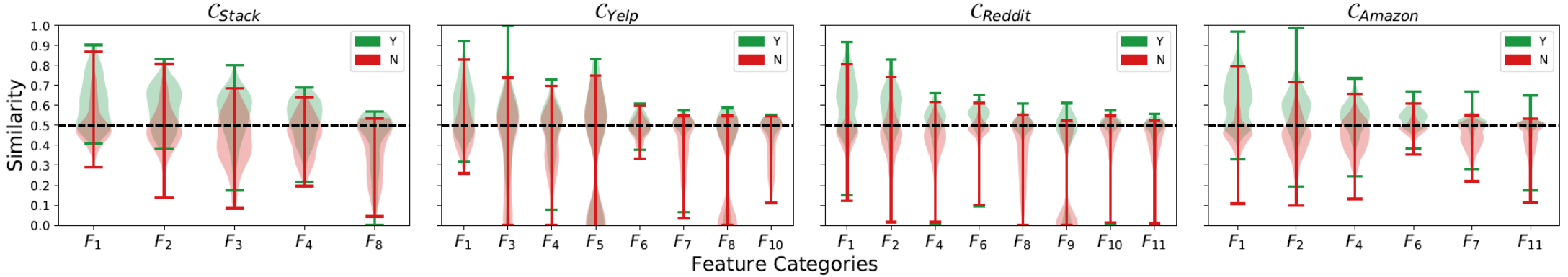
- To evaluate TAVeer, we made use of four self-compiled corpora (described in detail in the official paper) comprising verification cases with **cross**, **related** and **mixed** topics.
- Each corpus was partitioned into **author disjunct** training and test corpora based on a 40/60% ratio and was designed in a **balanced** manner (number of Y-cases equals the number of N-cases).
- In total, we have selected eight baseline methods (including the SOTA) that have shown their strengths in previous AV studies
- After training TAVeer and the respective baselines, we evaluated all methods on the four test corpora, using accuracy as a primary performance measure
- In two cases, TAVeer outperformed all baselines, while regarding the other two corpora it performed similar to the strongest baseline

	Method	Acc.	AUC	TP	FN	FP	TN
C_{Stack}	TAVeer	**0.697**	0.778	80	34	35	79
	IM	0.482	0.515	38	76	42	72
	BAFF	0.531	0.545	44	70	37	77
	DynamicAV	0.496	0.518	87	27	88	26
	NNCD	0.513	0.552	4	110	1	113
	ProfAV	0.539	0.609	67	47	58	56
	SPATIUM	0.636	0.723	49	65	18	96
	GenUnmasking	0.522	0.524	54	60	49	65
	Unmasking	0.539	0.542	60	54	51	63
C_{Yelp}	TAVeer	0.690	0.746	166	74	75	165
	IM	**0.708**	0.788	150	90	50	190
	BAFF	0.592	0.704	206	34	162	78
	DynamicAV	0.608	0.663	178	62	126	114
	NNCD	0.629	0.986	62	178	0	240
	ProfAV	0.665	0.723	155	85	76	164
	SPATIUM	0.590	0.651	93	147	50	190
	GenUnmasking	0.500	0.500	0	240	0	240
	Unmasking	0.596	0.639	153	87	107	133
C_{Reddit}	TAVeer	0.806	0.861	455	145	88	512
	IM	**0.833**	0.888	431	169	31	569
	BAFF	0.759	0.824	422	178	111	489
	DynamicAV	0.770	0.820	511	89	187	413
	NNCD	0.773	0.999	328	272	0	600
	ProfAV	0.764	0.821	453	147	136	464
	SPATIUM	0.797	0.863	446	154	90	510
	GenUnmasking	0.585	0.621	328	272	226	374
	Unmasking	0.719	0.785	467	133	204	396
C_{Amazon}	TAVeer	**0.842**	0.912	982	218	161	1039
	IM	0.815	0.901	941	259	186	1014
	BAFF	0.698	0.762	647	553	171	1029
	DynamicAV	0.785	0.876	1030	170	346	854
	NNCD	0.600	0.996	239	961	0	1200
	ProfAV	0.723	0.797	861	339	326	874
	SPATIUM	0.788	0.873	841	359	150	1050
	GenUnmasking	0.563	0.598	662	538	511	689
	Unmasking	0.725	0.801	903	297	362	838

Evaluation (Model analysis)

- To gain an insight into how the individual feature categories performed on the test corpora, we analyzed the trained models

- Using $\mathrm{sim}_{\mathfrak{C}}(\cdot)$, we calculated for all verification cases the aggregated similarity values with respect to the involved atomic ensembles in each model and visualized them as violin plots…



- **Interpretation:** The distribution of the similarity values for each F_i are colored green (**Y**) and red (**N**), respectively, while the dashed line represents the decision boundary. The better this line can separate both distributions and the less they overlap, the more suitable is F_i for the test corpus

Conclusions and Future Work

- To conclude our work, we would like to highlight the main characteristics of TAVeer:
  1) **Generalizability:** TAVeer can be effectively applied to verification cases with cross, related and mixed topics
  2) **Interpretability:** Using a simple scheme (described in our paper) one can interpret the verification results of the method, to gain insight into which specific features contributed to TAVeer's decision
  3) **Transparency:** All underlying text units (punctuations, words and phrases) used by TAVeer are predefined
  4) **Cross-domain ability:** TAVeer performs well across different domains

Corpus	Method	Model	Accuracy	AUC	TP	FN	FP	TN	N
Reddit [Test]	TAVeer	Reddit [Training]	0,806	0,861	455	145	88	512	1200
Reddit [Test]	TAVeer	Amazon [Training]	0,801	0,863	462	138	101	499	1200
Amazon [Test]	TAVeer	Amazon [Training]	0,842	0,912	982	218	161	1039	2400
Amazon [Test]	TAVeer	Reddit [Training]	0,810	0,900	959	241	216	984	2400

(Ongoing work...)

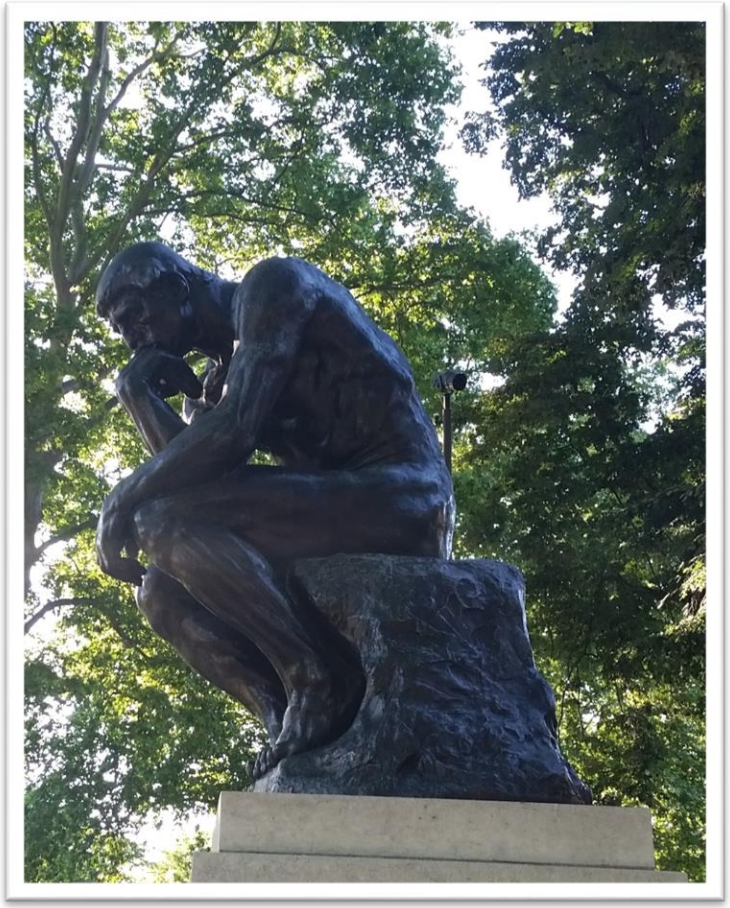
- **Directions for feature work:**
  - Currently, TAVeer cannot handle spelling mistakes. Therefore, we plan to counteract this by matching subword units rather than whole words. Moreover, we plan to investigate other feature categories including syntactic categories, abbreviations and interjections (e.g. "lol", "aha" or "hey")

Thank you very much for watching and listening…

Official paper: https://dl.acm.org/doi/10.1145/3407023.3409194
(Best paper award @ ARES/WSDF 2020)

Extended version @ arXiv: https://arxiv.org/abs/2006.12418

Talk @ YouTube: https://www.youtube.com/watch?v=hukRf40lp3g



Halvani, Oren. "***The Thinker***". 2017. Photograph.
Benjamin Franklin Parkway and 22nd Street, Philadelphia, USA.