Explainable Authorship Verification in Social Media via Attention-based Similarity Learning

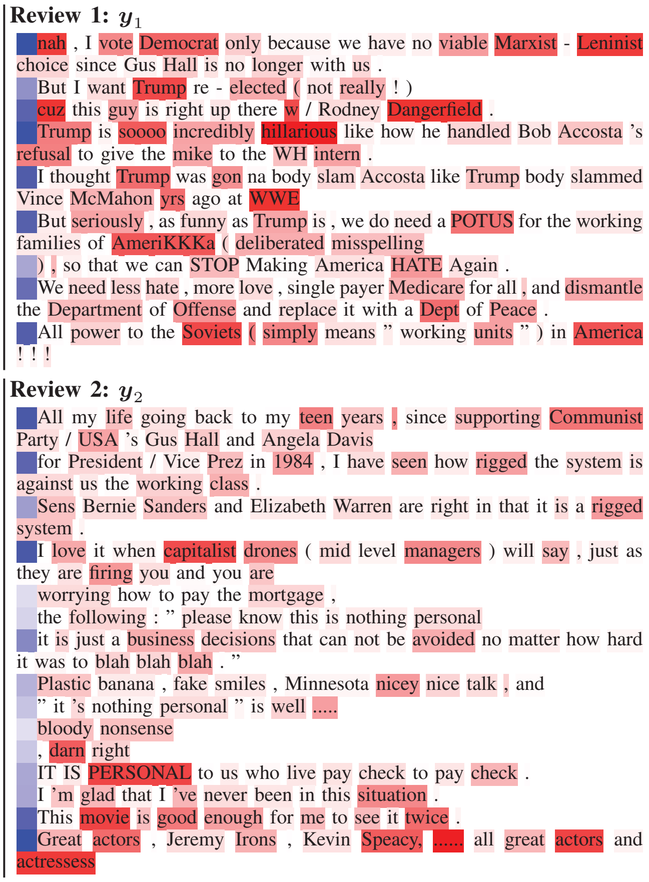
Benedikt Boenninghoff [1], Steffen Hessler[2], Dorothea Kolossa[1], Robert M. Nickel[3]
[1]*Cognitive Signal Processing Group, Ruhr University Bochum, Germany*
[2]*Department of German Philology , Ruhr University Bochum, Germany*
[3]*Department of Electrical and Computer Engineering, Bucknell University, Lewisburg, PA, USA*
Emails: benedikt.boenninghoff@rub.de, steffen.hessler@rub.de, dorothea.kolossa@rub.de, rmn009@bucknell.edu

***Abstract*—Authorship verification is the task of analyzing the linguistic patterns of two or more texts to determine whether they were written by the same author or not. The analysis is traditionally performed by experts who consider *linguistic* features, which include spelling mistakes, grammatical inconsistencies, and stylistics for example. Machine learning algorithms, on the other hand, can be trained to accomplish the same, but have traditionally relied on so-called *stylometric* features. The disadvantage of such features is that their reliability is greatly diminished for short and topically varied social media texts. In this interdisciplinary work, we propose a substantial extension of a recently published hierarchical Siamese neural network approach, with which it is feasible to learn *neural* features and to visualize the decision-making process. For this purpose, a new large-scale corpus of short Amazon reviews for text comparison research is compiled and we show that the Siamese network topologies outperform state-of-the-art approaches that were built up on stylometric features. Our linguistic analysis of the internal attention weights of the network shows that the proposed method is indeed able to latch on to some traditional linguistic categories.**

***Index Terms*—Authorship verification, similarity learning, forensic text comparison, Siamese network, deep metric learning**

I. Introduction

In authorship verification, also known as (forensic) text comparison, two or more text documents are compared with respect to their style to ascertain if the documents were written by the same author or not. Authorship verification is typically performed by experts who rely on traditional linguistic categories in their analysis. These categories include peculiarities of spelling/grammar, stylistic mannerisms, dialects, sociolects, and registers of language that hint at the authorship of a disputed document. Even though there is no such thing as a "linguistic fingerprint," the linguistic features people exhibit in their writing are specific enough to be admitted as evidence in court. The Federal Criminal Police (Bundeskriminalamt) in Germany, for example, uses text comparison techniques to identify potential suspects and/or to substantiate charges against a suspect in a criminal case [1]. A comprehensive overview of commonly employed categories in forensic linguistic can be found in [2]. Important practical implications of forensic linguistic, such as the relationship between language, criminal justice, and the law are explored in [3]. Research results on so-called idiolectal linguistic features that have become highly influential to authorship analysis and verification in general are reported in [4]. A particularly insightful description of the particular language used in fraud cases was provided in [5]. In addition, authorship verification is of great interest not only for the collection of evidence in criminal investigations, but also for the detection of deceptive intent and fake news in e-commerce and social media.



Review 1: y_1

nah , I vote Democrat only because we have no viable Marxist - Leninist choice since Gus Hall is no longer with us .
But I want Trump re - elected (not really !)
cuz this guy is right up there w / Rodney Dangerfield .
Trump is soooo incredibly hillarious like how he handled Bob Accosta 's refusal to give the mike to the WH intern .
I thought Trump was gon na body slam Accosta like Trump body slammed Vince McMahon yrs ago at WWE
But seriously , as funny as Trump is , we do need a POTUS for the working families of AmeriKKKa (deliberated misspelling
) , so that we can STOP Making America HATE Again .
We need less hate , more love , single payer Medicare for all , and dismantle the Department of Offense and replace it with a Dept of Peace .
All power to the Soviets (simply means " working units ") in America ! ! !

Review 2: y_2

All my life going back to my teen years , since supporting Communist Party / USA 's Gus Hall and Angela Davis
for President / Vice Prez in 1984 , I have seen how rigged the system is against us the working class .
Sens Bernie Sanders and Elizabeth Warren are right in that it is a rigged system .
I love it when capitalist drones (mid level managers) will say , just as they are firing you and you are
worrying how to pay the mortgage ,
the following : " please know this is nothing personal
it is just a business decisions that can not be avoided no matter how hard it was to blah blah blah . "
Plastic banana , fake smiles , Minnesota nicey nice talk , and
" it 's nothing personal " is well .....
bloody nonsense
, darn right
IT IS PERSONAL to us who live pay check to pay check .
I 'm glad that I 've never been in this situation .
This movie is good enough for me to see it twice .
Great actors , Jeremy Irons , Kevin Speacy, ... all great actors and actressess

Fig. 1: Attention-heatmap of our proposed authorship verification framework. Blue hues encode the sentence-based attention weights and red hues denote the relative word importance. All tokens are delimited by whitespaces.

Especially social media platforms have become an ubiquitous way of communication. Unfortunately, these platforms are also notorious for the proliferation of information from

This work was in significant parts performed on a high performance computing cluster at Bucknell University through the support of the National Science Foundation, Grant Number 1659397.

978-1-7281-0858-2/19/$31.00 ©2019 IEEE

unverified sources. As a result, users can fall victim to criminal predators through misinformation, fraud, and identity theft. The verification of the authorship of a piece of information can help to reduce the threat. The data volume shared on social media platforms on a daily basis, however, makes it utterly infeasible to rely on trained linguists for the analysis. Engineers and computer scientists have thus begun to automate parts of this process [6].

From a technical point of view, we can roughly distinguish between automatic authorship attribution and automatic authorship verification. The term *authorship attribution* describes a traditional classification task. Given a finite set of candidate authors, the objective is to determine who, from a set of enrolled authors, has written a document of unknown authorship [7]. In contrast, the objective of *authorship verification* or *text comparison* is to determine whether two separate documents were written by the same author [8], [9].

Methods for authorship analysis have traditionally been based on the extraction of stylometric features [10]–[16]. Stylometric features can generally be categorized into several distinct groups, e.g. lexical features, character features, syntactic features, semantic features, and application-specific features or compression-based features [17]–[19].

In contrast to stylometric-feature-based systems, there have also been a number of relatively recent papers that integrate the feature extraction task into a deep learning framework for authorship attribution [20], [21] and verification [22], [23].

With the advent of machine learning techniques, great strides have been made in the area of authorship verification by machine. The analysis of social media texts, however, still remains challenging [24]. Social media texts are often short, with a high variability in genre and topical content. The general difficulties that arise in authorship verification for social media are best illustrated with an example from the dataset that we are considering in Section III. Fig. 1 shows two product reviews from the Amazon e-commerce site. If we ignore the color coding for now (which will be explained later), we may quickly suspect that both reviews were written by the same author. Most prominently, we may perceive the particularly strong political stance implied in both texts. In addition, we may pick up on a few repetitive patterns. For instance, the author introduces a particular catch phrase and immediately follows it up with an "explanation" in parentheses, i.e. *"I want Trump re-elected (not really !)"* in the first review and *"capitalist drones (mid level managers)"* in the second review. Characteristic, also, is that the writing does not adhere to strict grammatical and spelling conventions (e. g. *"cuz"* used in the first review, third line in Fig. 2, which is an alternatively spelled abbreviation: *because* → *cause* → *cuz*) but rather violates these rules in very idiosyncratic ways, which makes it quite difficult to rely on part-of-speech tagging for analysis [25], for example. In addition, we may note that both texts exhibit strong similarities even though there is only a very limited overlap in the employed vocabulary. All of these aspects are readily noticeable for human beings, but imply significant challenges for machines. Earlier approaches have therefore had only limited success with social media data [26].

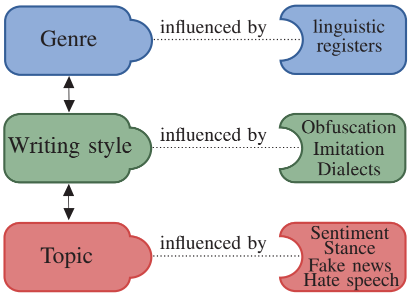
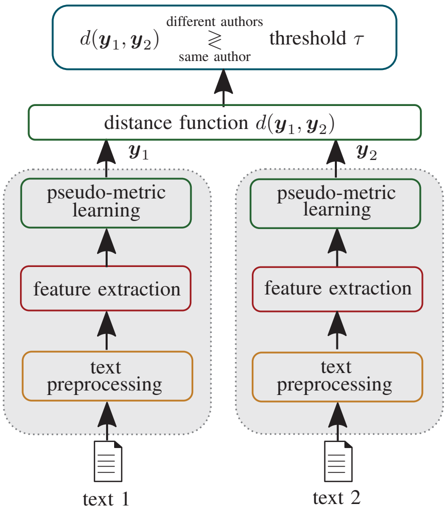


Fig. 2: Influences for an author's writing style.

An additional problem that affects authorship verification algorithms is that authors tend to dramatically shift the characteristics of their writing according to the situation they are in and according to the purpose they are envisioning for the text. Three specific examples are illustrated in Fig. 2:

- The *writing style* can be characterized by deviations from the standard writing style of a language. Stylistic variations may appear on all branches of linguistics (e.g. syntax, morphology, phonology, lexis, semantics). In order to automatically quantify deviations from the standard language, the text collections to be examined must be sufficiently long. The writing style can be influenced, for example, by imitation or obfuscation strategies.
- The writing style is generally also affected by the form of the text, i.e. whether we are dealing with a blackmail note, an Amazon review, a Tweet or a WhatsApp message. It has become customary, at least in parts of the literature, to refer to the type of a document generically as its *genre*. People change their linguistic register depending on the genre that they write in. This, in turn, leads to significant changes in computer linguistic characteristics of the resulting text. For a technical system it is thus extremely difficult to establish a common authorship between a WhatsApp message and a formal job application for example. It is therefore important to train classifiers only on one genre at a time.
- The *topic* is defined by the content of a text or the message that a person tries to communicate to the reader. The vocabulary that is used in a text tends to be strongly determined by the topic. Consequently, when the topic changes, then we observe a commensurate change in the vocabulary and thereby also a commensurate change in the characteristics of the text. It is thus desirable to develop authorship verification systems that do not put too much weight on similarity in vocabulary if we are dealing with cross-topic texts.

In many cases only very little text is (cumulatively) available from a respective author, which poses a great challenge for the training of any deep machine learning tool. In addition, machine learning tools that are trained in an unrestrained fashion tend to favor topical similarity between two texts over authorship similarity. This can lead to misclassifications if two texts from the same author treat different topics, or if two documents from different authors treat the same topic.



$d(\boldsymbol{y}_1, \boldsymbol{y}_2) \underset{\text{same author}}{\overset{\text{different authors}}{\gtrless}}$ threshold τ
distance function $d(\boldsymbol{y}_1, \boldsymbol{y}_2)$
$\boldsymbol{y}_1$
$\boldsymbol{y}_2$
pseudo-metric learning
pseudo-metric learning
feature extraction
feature extraction
text preprocessing
text preprocessing
text 1
text 2

Fig. 3: Overview of the proposed method ADHOMINEM.

Motivated by this fact, we propose a novel neural network topology for the text comparison task that is applicable to a big data collection of social media texts. The technical core of our approach is implemented by an *Attention-based Deep Hierarchical cOnvolutional siaMese bIdirectional recurreNt nEural-network Model* (ADHOMINEM).

We specifically propose a substantial extension of our previous work presented in [27], where a hierarchical recurrent Siamese network (HRSN) was applied to encode an entire document into a single vector. Authorship analysis generally is an interdisciplinary field in which it is not only important to establish *who* the suspected author is, but equally important to establish *how* this decision was arrived at. The question of *how* is crucial if we should ever hope to have neural network-based methods stand up in court. From this aspect, our previous framework (HRSN) has two drawbacks: It does not use linguistically interpretable stylometric features, and it is not able to directly visualize the decision-making process. In contrast, ADHOMINEM has the capacity to automatically learn meaningful neural features from a big data corpus that latch on to some traditional linguistic categories.

As illustrated in Fig. 3, our ADHOMINEM topology includes three important stages: The first stage contains a text preprocessing step. The second stage includes a feature extraction in which we encode each document consisting of characters, words, and sentences into one single *neural feature vector*, denoted by $\boldsymbol{y}_i$ for $i \in 1, 2$ [28]. We incorporate a characters-to-word encoding layer [29] to take the specific uses of prefixes and suffixes as well as spelling errors into account. Additionally, an incorporation of attention layers [30] allows us to visualize words and sentences that have been marked as *highly significant* by the system. In the third stage, we employ a module for nonlinear metric learning to measure the similarity between two documents [31].

We, furthermore, defined/constructed a new large dataset based on Amazon reviews for the study of authorship verification tasks. Amazon reviews do not represent forensic texts from a law enforcement/criminal point of view. We argue, nevertheless, that one can still gain valuable insights from the data into how to accomplish forensic text analysis in general. The advantages of the defined dataset are:

- A very large amount of reviews is publicly available and, to our best knowledge, there exists no comparable large-scale dataset containing forensic texts.
- Social media texts usually contain many *"easy-to-visualize"* linguistic features such as spelling errors, which are crucial for forensic text comparison in general.
- The provided social media corpus is annotated w.r.t authorship and topical category. Hence, by fixing the genre (see Fig. 2), we are able to analyze to which degree the authorship decision is influenced by the topic.
- For Amazon reviews, the number of authors trying to obfuscate their authorship is vanishingly small and obfuscation is, thus, not likely to bias our analysis of the context sensitivity.
- With our proposed model, it is straightforward to incorporate a domain adaptation using smaller-sized forensic datasets, so we can easily adapt ADHOMINEM to forensic text comparison.

In summary, this paper provides the following contributions:

1) We propose a novel attention-based Siamese network topology with applications in authorship verification for large social media datasets.
2) We have prepared a new large-scale corpus of short Amazon reviews for authorship verification research.
3) In addition to a quantitative evaluation of the proposed method we present a qualitative linguistic analysis of visualized attention weights and provide evidence that the visualized attentions can provide an explanation for the decision of the network.

II. ATTENTION-BASED SIAMESE NETWORK TOPOLOGY

With our ADHOMINEM approach we propose a significant extension to our model introduced in [27]. Its Siamese topology consists of two identical neural networks that share the exact same set of parameters. The network is trained to extract document-specific features to make a similarity analysis between two documents as reliable and robust as possible. The overall architecture is depicted in Fig. 4. Stage one, i.e. the preprocessing of the input texts, is described in Section IV-A. Details of stages two and three are discussed below.

A. Characters-to-word Encoding

Let $\boldsymbol{x}^{(\mathrm{c})}_{i,j,k}$ be the real-valued D_{c}-dimensional character embedding vector that corresponds to the i-th character of the j-th word in the k-th sentence. We concatenate h character embeddings to form a $D_{\mathrm{c}} \cdot h$-dimensional vector $\boldsymbol{x}^{(\mathrm{c})}_{i:i+h-1,j,k} = \boldsymbol{x}^{(\mathrm{c})}_{i,j,k} \oplus \boldsymbol{x}^{(\mathrm{c})}_{i+1,j,k} \oplus \ldots \oplus \boldsymbol{x}^{(\mathrm{c})}_{i+h-1,j,k}$ and apply one-dimensional convolution,

$$\boldsymbol{c}_{i,j,k} = \tanh\big(\boldsymbol{W}^{(\mathrm{c})} \boldsymbol{x}^{(\mathrm{c})}_{i:i+h-1,j,k} + \boldsymbol{b}^{(\mathrm{c})}\big), \qquad (1)$$

where $\oplus$ defines the concatenation operator [32]. In Eq. (1), the set $\boldsymbol{\theta}^{(\mathrm{CNN})} = \{\boldsymbol{W}^{(\mathrm{c})} \in \mathbb{R}^{D_{\mathrm{r}} \times h \cdot D_{\mathrm{c}}}, \, \boldsymbol{b}^{(\mathrm{c})} \in \mathbb{R}^{D_{\mathrm{r}} \times 1}\}$ represents trainable parameters. Applying max-over-time pooling w.r.t. all D_r-dimensional vectors $\boldsymbol{c}_{i,j,k}$ results in

$$\boldsymbol{r}_{j,k} = \max_{1 \leq i \leq T^{(\mathrm{s})}-h+1} \{\boldsymbol{c}_{i,j,k}\}, \tag{2}$$

where $\boldsymbol{r}_{j,k} \in \mathbb{R}^{D_\mathrm{r} \times 1}$ denotes the *character representation* of the j-th word in the k-th sentence [29].

B. Words-to-sentence Encoding

Let $\boldsymbol{x}^{(\mathrm{w})}_{j,k} \in \mathbb{R}^{D_\mathrm{w} \times 1}$ be the word embedding of the j-th word in the k-th sentence. We now feed concatenations of character and word representations into the bidirectional LSTM network. The forward path at time $j \in \{1, \dots, T^{(\mathrm{w})}\}$ can be written as

$$\big\{\overrightarrow{\boldsymbol{h}}^{(\mathrm{w})}_{j,k}, \overrightarrow{\boldsymbol{c}}^{(\mathrm{w})}_{j,k}\big\} = \mathrm{LSTM}_{\boldsymbol{\theta}^{(\mathrm{fRNN})}_{\mathrm{ws}}}\left(\begin{bmatrix}\boldsymbol{x}^{(\mathrm{w})}_{j,k} \\ \boldsymbol{r}_{j,k}\end{bmatrix}, \overrightarrow{\boldsymbol{h}}^{(\mathrm{w})}_{j-1,k}, \overrightarrow{\boldsymbol{c}}^{(\mathrm{w})}_{j-1,k}\right), \tag{3}$$

where the hidden state and the memory state are denoted by $\overrightarrow{\boldsymbol{h}}^{(\mathrm{w})}_{j,k} \in \mathbb{R}^{D_\mathrm{s} \times 1}$ and $\overrightarrow{\boldsymbol{c}}^{(\mathrm{w})}_{j,k} \in \mathbb{R}^{D_\mathrm{s} \times 1}$, respectively. The set $\boldsymbol{\theta}^{(\mathrm{fRNN})}_{\mathrm{ws}}$ is comprised of all trainable parameters of the forward LSTM cell. Analogously, we denote the parameter set of the backward path with $\boldsymbol{\theta}^{(\mathrm{bRNN})}_{\mathrm{ws}}$. The joint hidden state is given by the concatenation of the forward and backward states, i.e. $\boldsymbol{h}^{(\mathrm{w})}_{j,k} = \overrightarrow{\boldsymbol{h}}^{(\mathrm{w})}_{j,k} \oplus \overleftarrow{\boldsymbol{h}}^{(\mathrm{w})}_{j,k}$. According to [28] we incorporate an attention layer in the form

$$\begin{aligned} \alpha^{(\mathrm{w})}_{j,k} &= \frac{\exp\{\boldsymbol{v}^{(\mathrm{a})}_{\mathrm{ws}} \boldsymbol{u}^{(\mathrm{w})}_{j,k}\}}{\sum_{j'=1}^{T^{(\mathrm{w})}} \exp\{\boldsymbol{v}^{(\mathrm{a})}_{\mathrm{ws}} \boldsymbol{u}^{(\mathrm{w})}_{j',k}\}} \\ \boldsymbol{x}^{(\mathrm{s})}_{k} &= \sum_{j=1}^{T^{(\mathrm{w})}} \alpha^{(\mathrm{w})}_{j,k} \boldsymbol{h}^{(\mathrm{w})}_{j,k}, \end{aligned} \tag{4}$$

where $\boldsymbol{u}^{(\mathrm{w})}_{j,k} = \tanh\big(\boldsymbol{W}^{(\mathrm{a})}_{\mathrm{ws}} \boldsymbol{h}^{(\mathrm{w})}_{j,k} + \boldsymbol{b}^{(\mathrm{a})}_{\mathrm{ws}}\big)$ and $\boldsymbol{x}^{(\mathrm{s})}_{k}$ denotes the k-th *sentence embedding*. Trainable parameters are $\boldsymbol{\theta}^{(\mathrm{ATT})}_{\mathrm{ws}} = \{\boldsymbol{W}^{(\mathrm{a})}_{\mathrm{ws}} \in \mathbb{R}^{D^{(\mathrm{a})}_{\mathrm{ws}} \times 2 \cdot D_\mathrm{s}}, \boldsymbol{b}^{(\mathrm{a})}_{\mathrm{ws}} \in \mathbb{R}^{D^{(\mathrm{a})}_{\mathrm{ws}} \times 1}, \boldsymbol{v}^{(\mathrm{a})}_{\mathrm{ws}} \in \mathbb{R}^{1 \times D^{(\mathrm{a})}_{\mathrm{ws}}}\}$.

C. Sentences-to-document Encoding

On the next tier, we feed the obtained sentence embeddings into another bidirectional LSTM cell. The forward path at the k-th time step with $k \in \{1, \dots, T^{(\mathrm{s})}\}$ is given by

$$\big(\overrightarrow{\boldsymbol{h}}^{(\mathrm{s})}_{k}, \overrightarrow{\boldsymbol{c}}^{(\mathrm{s})}_{k}\big) = \mathrm{LSTM}_{\boldsymbol{\theta}^{(\mathrm{fRNN})}_{\mathrm{sd}}}\big(\boldsymbol{x}^{(\mathrm{s})}_{k}, \overrightarrow{\boldsymbol{h}}^{(\mathrm{s})}_{k-1}, \overrightarrow{\boldsymbol{c}}^{(\mathrm{s})}_{k-1}\big), \tag{5}$$

where sentence-based hidden and memory states are given by $\overrightarrow{\boldsymbol{h}}^{(\mathrm{s})}_{k} \in \mathbb{R}^{D_\mathrm{d} \times 1}$ and $\overrightarrow{\boldsymbol{c}}^{(\mathrm{s})}_{k} \in \mathbb{R}^{D_\mathrm{d} \times 1}$. The set $\boldsymbol{\theta}^{(\mathrm{fRNN})}_{\mathrm{sd}}$ is comprised of all trainable parameters of the forward LSTM cell. Again, we denote the parameter set of the backward path with $\boldsymbol{\theta}^{(\mathrm{bRNN})}_{\mathrm{sd}}$ and the combined/joint hidden state with $\boldsymbol{h}^{(\mathrm{s})}_{k} = \overrightarrow{\boldsymbol{h}}^{(\mathrm{s})}_{k} \oplus \overleftarrow{\boldsymbol{h}}^{(\mathrm{s})}_{k}$. The attention layer at the sentence-level is then defined analogously to (4),

$$\begin{aligned} \alpha^{(\mathrm{s})}_{k} &= \frac{\exp\{\boldsymbol{v}^{(\mathrm{a})}_{\mathrm{sd}} \boldsymbol{u}^{(\mathrm{s})}_{k}\}}{\sum_{k'=1}^{T^{(\mathrm{w})}} \exp\{\boldsymbol{v}^{(\mathrm{a})}_{\mathrm{sd}} \boldsymbol{u}^{(\mathrm{s})}_{k'}\}} \\ \boldsymbol{x}^{(\mathrm{d})} &= \sum_{k=1}^{T^{(\mathrm{s})}} \alpha^{(\mathrm{s})}_{k} \boldsymbol{h}^{(\mathrm{s})}_{k}, \end{aligned} \tag{6}$$

with $\boldsymbol{u}^{(\mathrm{s})}_{k} = \tanh\big(\boldsymbol{W}^{(\mathrm{a})}_{\mathrm{sd}} \boldsymbol{h}^{(\mathrm{s})}_{k} + \boldsymbol{b}^{(\mathrm{a})}_{\mathrm{sd}}\big)$ and $\boldsymbol{x}^{(\mathrm{d})}$ representing the *document embeddings*. Trainable parameters are $\boldsymbol{\theta}^{(\mathrm{ATT})}_{\mathrm{sd}} = \{\boldsymbol{W}^{(\mathrm{a})}_{\mathrm{sd}} \in \mathbb{R}^{D^{(\mathrm{a})}_{\mathrm{sd}} \times 2 \cdot D_\mathrm{d}}, \boldsymbol{b}^{(\mathrm{a})}_{\mathrm{sd}} \in \mathbb{R}^{D^{(\mathrm{a})}_{\mathrm{sd}} \times 1}, \boldsymbol{v}^{(\mathrm{a})}_{\mathrm{sd}} \in \mathbb{R}^{1 \times D^{(\mathrm{a})}_{\mathrm{sd}}}\}$.

D. Nonlinear Metric Learning

We now feed the document embbedings $\boldsymbol{x}^{(\mathrm{d})}$ into a fully-connected multilayer perceptron (MLP),

$$\boldsymbol{y} = \tanh\big(\boldsymbol{W}^{^{(\mathrm{f})}} \boldsymbol{x}^{(\mathrm{d})} + \boldsymbol{b}^{^{(\mathrm{f})}}\big), \tag{7}$$

to obtain the D_f-dimensional *document features* $\boldsymbol{y}$ [33]. The parameter set associated with the MLP is denoted with $\boldsymbol{\theta}^{(\mathrm{MLP})}$ = $\{\boldsymbol{W}^{^{(\mathrm{f})}} \in \mathbb{R}^{D_\mathrm{f} \times 2 \cdot D_\mathrm{d}}, \boldsymbol{b}^{(\mathrm{f})} \in \mathbb{R}^{D_\mathrm{f} \times 1}\}$. Given a pair of document features, $\boldsymbol{y}_i$ for $i \in \{1, 2\}$ via Eq. (7), we can measure the similarity of both documents by determining the *Euclidean* distance,

$$d\big(\boldsymbol{y}_1, \boldsymbol{y}_2\big) = \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2 \,. \tag{8}$$

E. The Loss Function

The entire network is trained end-to-end, such that documents written by the same author should result in small values for Eq. (8), while the measure should return large values for documents of different authors. A key problem in automatic text comparison is that it is generally much easier to compare the topical content of two documents than it is to compare their authorship. It is, thus, important that feature vectors in Eq. (7) are made insensitive to topical variations between texts. A practical means to accomplish this insensitivity has been proposed in [27], [33]. The approach is based on a *double threshold* concept illustrated in the top left corner of Figure 4. Two distance thresholds τ_s and τ_d are defined with $\tau_s < \tau_d$. During training, all distances between document pairs (🟥,🟥) that are considered to belong to a same-author category (labeled with $a = 1$) are to stay below the lower of the two thresholds, τ_s, e.g.

$$\mathcal{L}^{(\mathrm{s})}_{\boldsymbol{\Theta}} = a \cdot \max\big\{d\big(\boldsymbol{y}_1, \boldsymbol{y}_2\big) - \tau_s, 0\big\}^2, \tag{9}$$

with all trainable parameters $\boldsymbol{\Theta} = \{\boldsymbol{\theta}^{(\mathrm{CNN})}, \boldsymbol{\theta}^{(\mathrm{fRNN})}_{\mathrm{ws}}, \boldsymbol{\theta}^{(\mathrm{bRNN})}_{\mathrm{ws}}, \boldsymbol{\theta}^{(\mathrm{ATT})}_{\mathrm{ws}}, \boldsymbol{\theta}^{(\mathrm{fRNN})}_{\mathrm{sd}}, \boldsymbol{\theta}^{(\mathrm{bRNN})}_{\mathrm{sd}}, \boldsymbol{\theta}^{(\mathrm{ATT})}_{\mathrm{sd}}, \boldsymbol{\theta}^{(\mathrm{MLP})}\}$. Conversely, distances between document pairs (🟥,▲) that belong to the different-author category (labeled with $a = 0$) are to remain above the higher threshold τ_d, e.g.

$$\mathcal{L}^{(\mathrm{d})}_{\boldsymbol{\Theta}} = (1 - a) \cdot \max\big\{\tau_d - d\big(\boldsymbol{y}_1, \boldsymbol{y}_2\big), 0\big\}^2. \tag{10}$$

The final loss function is then given by

$$\mathcal{L}_{\boldsymbol{\Theta}} = \mathcal{L}^{(\mathrm{s})}_{\boldsymbol{\Theta}} + \mathcal{L}^{(\mathrm{d})}_{\boldsymbol{\Theta}}. \tag{11}$$

Note that the loss function in Eq. (11) itself does not directly deal with the problem of topical variations. Our motivation behind the double-threshold-mechanism is that algorithms generally tend to misclassify short texts annotated with *same-author/cross-topics* or *different-authors/same-topic*. We therefore suggest to combine the loss function in Eq. (11) with a large, balanced dataset, where we have the *same* number of occurrences of *same-topic* and *cross-topic* cases to deal with. Document pairs that are easy to verify (*same-author/same-topic* and *different-author/cross-topic*) are ignored during training when their distances are under ($\leq \tau_s$) or above ($\geq \tau_d$) the corresponding threshold. As a result, our system strongly focuses on more difficult document pairs.

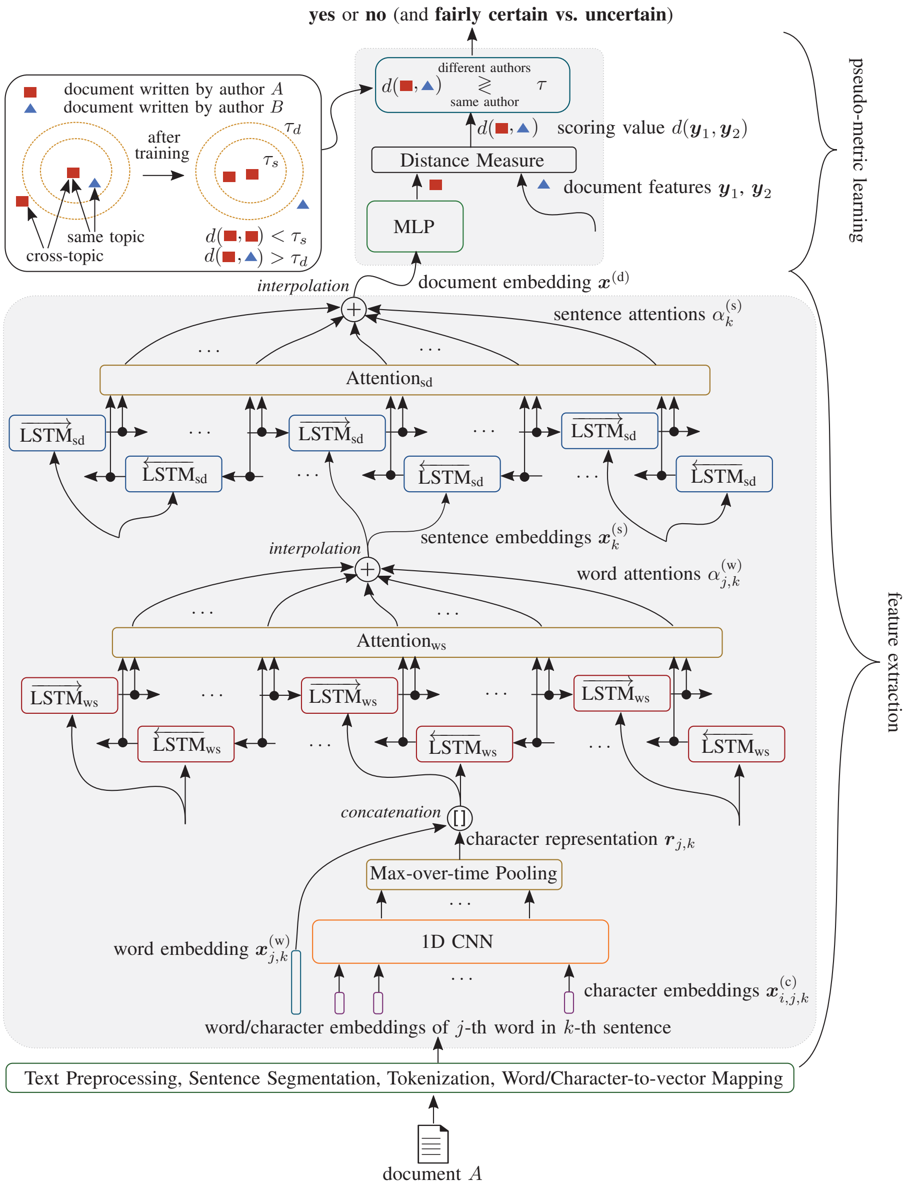

yes or no (and fairly certain vs. uncertain)
document written by author A
document written by author B
after training
τ_d
τ_s
same topic
cross-topic
$d(■, ■) < \tau_s$
$d(■, ▲) > \tau_d$
different authors
same author
$d(■, ▲) \gtrless \tau$
$d(■, ▲)$ scoring value $d(\boldsymbol{y}_1, \boldsymbol{y}_2)$
Distance Measure
document features $\boldsymbol{y}_1$, $\boldsymbol{y}_2$
MLP
pseudo-metric learning
interpolation
document embedding $\boldsymbol{x}^{(d)}$
sentence attentions $\alpha_k^{(s)}$
Attention$_{sd}$
$\overrightarrow{\text{LSTM}}_{sd}$
$\overleftarrow{\text{LSTM}}_{sd}$
interpolation
sentence embeddings $\boldsymbol{x}_k^{(s)}$
word attentions $\alpha_{j,k}^{(w)}$
Attention$_{ws}$
$\overrightarrow{\text{LSTM}}_{ws}$
$\overleftarrow{\text{LSTM}}_{ws}$
feature extraction
concatenation
character representation $\boldsymbol{r}_{j,k}$
Max-over-time Pooling
1D CNN
word embedding $\boldsymbol{x}_{j,k}^{(w)}$
character embeddings $\boldsymbol{x}_{i,j,k}^{(c)}$
word/character embeddings of j-th word in k-th sentence
Text Preprocessing, Sentence Segmentation, Tokenization, Word/Character-to-vector Mapping
document A

Fig. 4: Architecture of the proposed ADHOMINEM system. The feature extraction system for document B (blue triangle) is not visible, since it is identical with the feature extraction system for document A (red square).

III. LARGE-SCALE CORPUS FOR TEXT COMPARISON

We prepared a new large-scale corpus of short Amazon reviews as follows [34]:

- Preprocessing steps are applied as discussed in Section IV-A. All reviews with less then 80 tokens and more than 1000 tokens were removed. As a result, we obtained 9,052,606 reviews written by 784,649 authors, where, on average, each review consists of 282.10 ± 198.14 tokens. Additionally, the reviews for each author are grouped into 24 different categories (by their meta-data)[1] to be able to analyze cross-topic/same topic instances.
- We removed all rare token types with less than 20 overall occurrences to reduce the vocabulary size from 7,427,762 to 317,712 tokens. Analogously, we also removed all character types with less 100 overall occurrences to reduce the size from 1,482 to 222 characters.
- In addition to *zero-padding* tokens (for short sentences) and *unknown* tokens, we also introduced tokens to deal with long sentences. If a sentence is shorter than a predefined maximum sentence length then it ends with a regular *sentence-ends* token. If a sentence is longer than the maximum sentence length then we stop with a *line-break* token and shift the remaining part of the sentence into the next line.
- For cross-validation, the reviews are randomly split into disjoint groups w.r.t. the authorship. As a result, development and test sets only contain unseen reviews written by unseen authors.
- We sample review pairs w.r.t. to the authorship and cat-

[1] Raw dataset available at `http://jmcauley.ucsd.edu/data/amazon`

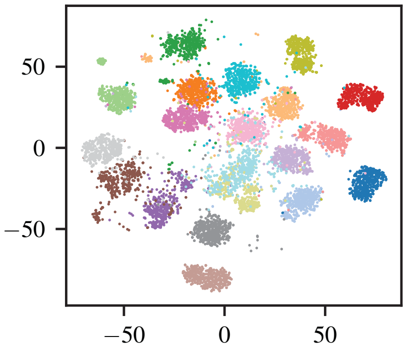


Fig. 5: A t-SNE-plot of neural features after [35]. We randomly selected 500 reviews written by 20 *unseen* authors. Each of the color belongs to another author. The representation illustrates the discriminative power of ADHOMINEM over unseen data.

egory. Each review pair is labeled by a tuple $l = (a, c)$ with $a \in \{0, 1\}$ and $c \in \{0, 1\}$. The value of a indicates if the reviews were written by the same author ($a = 1$) or by different authors ($a = 0$). The value of c indicates if the reviews treat the same topic ($c = 1$) or treat different topics ($c = 0$).

- The review pairs are recombined after each epoch in order to increase the heterogeneity of the training set.
- Each author contributes with only a minimum number of documents w.r.t. each tuple category $l = (a, c)$ which are represented equally to yield a balanced dataset.

Altogether we obtain around $335,000$ training pairs for each epoch and $42,000$ instances for the development/test sets each, with an equal number of instances for all labels $l = (a, c)$.

IV. EVALUATION

In this section, we present experimental results for the proposed ADHOMINEM system in comparison to a few other published approaches. A detailed documentation of the training procedure, including hyper-parameter settings, program code, and the data are accessible [2] for other researchers in the field.

A. Implementation Details

ADHOMINEM was implemented in Python. We utilized the library `textacy`[3] for data preprocessing. All URLs, email addresses, and phone numbers were replaced with respective universal tokens as they themselves are typically not part of an author's writing style but only the *use of them* is part of an author's writing style. The package `spaCy`[4] was used for sentence boundary detection and tokenization. The training of the neural networks was accomplished with `Tensorflow`[5]. Pretrained word embeddings were taken from `fastText` [36].

B. Baseline Methods

We chose four published authorship verification methods as comparison references for our proposed approach.

TABLE I: Average verification error rates and standard deviations for the test set over a 5-fold cross-validation.

labels	error rate in %				
	IMPOSTERS	AVEER	GLAD	HRSN	ADHOMINEM
$\forall l{=}(a,c)$	33.33 ± 0.16	30.13 ± 0.13	27.13 ± 0.14	15.40 ± 0.19	14.70 ± 0.16
$a{=}1,\ c{=}1$	25.76 ± 0.59	25.63 ± 0.25	24.27 ± 0.23	11.74 ± 0.12	11.13 ± 0.52
$a{=}1,\ c{=}0$	40.27 ± 1.13	35.88 ± 0.53	35.97 ± 0.57	17.81 ± 0.33	16.74 ± 0.75
$a{=}0,\ c{=}1$	42.54 ± 0.94	36.37 ± 0.54	30.28 ± 0.54	22.20 ± 0.65	21.24 ± 1.03
$a{=}0,\ c{=}0$	24.87 ± 1.21	22.64 ± 0.37	18.02 ± 0.41	9.85 ± 0.45	9.68 ± 0.79

TABLE II: Average verification error rates and standard deviations for the test set over a 5-fold cross-validation. Only results for which the systems reported a **high reliability** are considered.

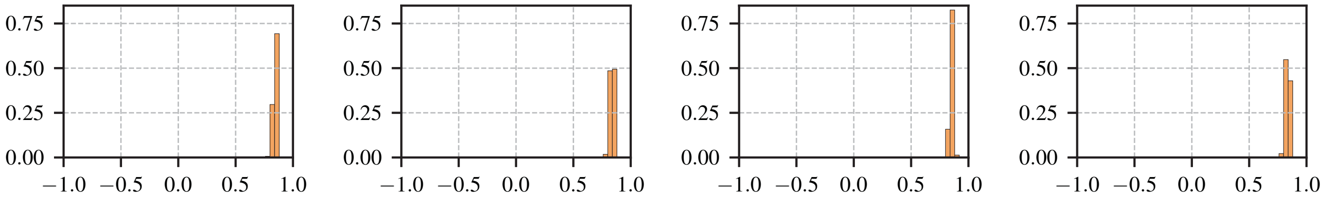
	$d(\boldsymbol{y}_1, \boldsymbol{y}_2) \leq \tau_s$ and $d(\boldsymbol{y}_1, \boldsymbol{y}_2) \geq \tau_d$			
labels	HRSN		ADHOMINEM	
	error rate (%)	# instances (%)	error rate (%)	# instances (%)
$\forall l{=}(a,c)$	0.93 ± 0.04	19.91 ± 0.49	0.84 ± 0.07	21.30 ± 0.41
$a{=}1,\ c{=}1$	1.15 ± 0.33	13.08 ± 0.38	1.05 ± 0.17	14.66 ± 0.90
$a{=}1,\ c{=}0$	3.12 ± 0.43	12.07 ± 0.61	2.74 ± 0.49	13.35 ± 0.55
$a{=}0,\ c{=}1$	0.99 ± 0.20	19.60 ± 0.60	0.81 ± 0.20	20.93 ± 1.18
$a{=}0,\ c{=}0$	0.07 ± 0.02	34.89 ± 1.06	0.09 ± 0.03	36.25 ± 1.38

AVEER [15], GLAD [37] and IMPOSTERS [9] are based on a traditional stylometric feature extraction. These three algorithms have been ranked first, second and third in a performance evaluation on a small-sized corpus of larger Amazon reviews conducted by [38]. In addition, we also considered our predecessor HRSN [27].

C. Results

Table I summarizes the average verification error rates for the dataset described in Section III with a 5-fold cross-validation. It is readily seen from the first row of Table I that our two methods based on a Siamese topology (HRSN and ADHOMINEM) significantly outperformed the other three systems which are based on traditional stylometric features. Comparing the results of the baseline methods IMPOSTERS, AVEER and GLAD, we obtained error rates between $27\% - 34\%$, while both, HRSN and ADHOMINEM, were able to cut the error rate in half to around $14\% - 16\%$. Comparing HRSN and ADHOMINEM, we were able to slightly increase the average accuracy with our new attention-based Siamese topology. Rows 2-5 additionally show the performance w.r.t. the different label categories defined in Section III. As it can be seen, all systems perform well for *same-author/same-category* instances as well as for *different-author/cross-topic* instances. As one would expect, the error rates for all methods dramatically increase for *same-author/cross-topic* and *different-author/same-topic* cases. The results presented in Table I show very clearly that the dataset discussed in Section III is quite challenging for all methods. The proposed Siamese network topologies, however, displayed a significantly higher discriminative power than the stylometric-feature-based systems. This is also evident in Fig. 5, where, in a first step, we randomly selected 500

[2] https://github.com/rub-ksv/AdHominem
[3] https://chartbeat-labs.github.io/textacy
[4] https://spacy.io/
[5] https://www.tensorflow.org/



(a) mean/std: 0.846 ± 0.015 (b) mean/std: 0.838 ± 0.017 (c) mean/std: 0.853 ± 0.014 (d) mean/std: 0.836 ± 0.016

Fig. 6: Histograms of Kendall-τ correlation coefficients between weighted attentions of the reference run and all other cross-validations. This correlation is statistically significant ($p \leq 0.01$) for all instances. Values near 0 indicate no correspondence and values close to 1 imply perfect agreement.

documents written by 20 different *unseen* authors; and, in a second step, we computed the neural features $\boldsymbol{y}$ using the trained feature extraction module of ADHOMINEM for each of the documents. In a third step, we applied a t-SNE to reduce the dimension to produce a visual representation. Each color in the plot belongs to a different author. As it can be seen, the features $\boldsymbol{y}$ produced by ADHOMINEM are well suited to discriminate between different *unseen* authors.

In addition to overall error rate counts, we are also reporting the degree to which the Siamese network topologies are able to decide with a *high level of reliability*. The *double threshold* concept introduced in Section II-E can be leveraged to this end. If the feature vector tuples $(\boldsymbol{y}_1, \boldsymbol{y}_2) =$ (🟥, 🟥) or $(\boldsymbol{y}_1, \boldsymbol{y}_2) =$ (🟥, ▲) of two texts have a distance $d(\boldsymbol{y}_1, \boldsymbol{y}_2)$ below τ_s or above τ_d then we may assume that the system attaches a high reliability to its decision. If the distance is between τ_s and τ_d, however, then we may still arrive at a decision by comparing the distance to $\frac{\tau_s+\tau_d}{2}$, but the decision would carry much less "confidence". Results for text comparisons for which a *high level of reliability* is reported are presented in Table II. We are separately reporting on the performance of ADHOMINEM and HRSN regarding cross-topic and same topic cases. In Table II, we observe that the error rates generally decrease dramatically when we only consider cases where the scores exceed our predefined thresholds. With both Siamese topologies we may reduce the error rate below 1%. It should be noted, however, that this holds for only a limited number of cases in our test set. The respective number of instances for which a *high reliability* is detected is reported in Table II as well. The overall error rate, as reported in the first row of Table II, is similar between ADHOMINEM and HRSN, but ADHOMINEM improves slightly in the number of instances in which this is the case.

D. Correlation Analysis

In Section IV-E we are providing a linguistic analysis of visualized attentions based on a single reference run from our 5-fold cross-validation. If the resulting attention weights bear merit as indicators of "linguistic importance" then we would expect that the weights obtained from different runs should be highly correlated. In order to study this correlation we formed an overall attention weight $\alpha_{j,k}$ as the product between the word-based attentions from Eq. (4) and the associated sentence-based attentions from Eq. (6), i.e.

$$\alpha_{j,k} = \alpha_{j,k}^{(\mathrm{w})} \cdot \alpha_k^{(\mathrm{s})}. \tag{12}$$

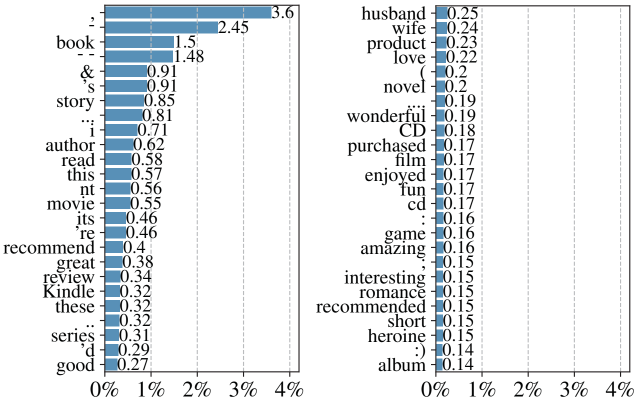


Fig. 7: Frequencies of all tokens being listed in the top 5 ranking w.r.t the attention weights in a review.

We determined Kendall-τ coefficients of the $\alpha_{j,k}$ between the reference and all other cross-validation runs [39]. The results are shown in Fig. 6. The weighted attentions $\alpha_{j,k}$ in Eq. (12) from different cross-validation runs exhibit a high degree of correlation with the reference run. This supports our claim that attention weights may be used as indicators of "linguistic importance".

E. Linguistic analysis

As already mentioned in the previous section, we were motivated by [40] to check if any of the ADHOMINEM attention weights would exhibit traits of linguistic categories. In order to support the analysis we utilize Eq. (12) to color-code the texts as shown below, i.e. a red background implies a high attention weight and a white background implies a low attention weight.

Fig. 7 provides a list of the collected top 5 ranked tokens of each document that were tallied/counted across all documents. At the top of the list we only see *known* tokens (for which we trained word embeddings), which is not surprising. Grammatical errors or mispellings are performed individually. However, our proposed system is able to handle *unknown* tokens, e.g. the following found misspellings of the token *definitely*:

> *diffenatly, definatley, definately, definitley, definitly, definetly, defintely, definatly, definalty, definantly, definetely, definitely(and, definently*

Cumulatively, they represent 0.13% of the top 5 tokens.

We examined 100 randomly chosen text comparisons and analyzed them in order to extract linguistic features. No linguistic category was excluded, but the analysis focused on those parts that had high attention weights. A limited number of examples of identified linguistic features, as highlighted by the system, is provided below.

Punctuation

1) Some punctuation marks (e.g. hyphen, brackets, colon, comma) are frequently marked:

> **Example 1: -**
> Clarke is just satisfactory - her Jane is too weak and bland .

2) Accumulation of punctuation marks:

> **Example 2: ...**
> I just read the chapter on Generosity ... and it was PHENOMENAL !

3) Special characters:

> **Example 3: &**
> [. . .] if you need to tweak & bend them

4) Missing white spaces:

> **Example 4: book.5 vs. books. 5**
> I highly recommend this book.5 Stars

Characters

5) Substitutions of characters (typing errors):

> **Example 5: deffinatly vs. definitely**
> Ice Cube deffinatly has a style that is unimatatable .

6) Missing characters:

> **Example 6: clasics vs. classics**
> [. . .] he went on to put out some hip hop clasics .

7) Surplus (redundant) characters:

> **Example 7: amazone vs. Amazon/amazon**
> i searched a lot of this kind of high tech light around amazone

Capitalization

8) Lower instead of upper case:

> **Example 8: i vs. I**
> [. . .] i 'm glad i picked this .

9) Upper instead of lower case:

> **Example 9: Stars vs. stars**
> I highly recommend this book.5 Stars

10) Continuous capitalization:

> **Example 10: TOTALLY AWESOME**
> Uncaged is TOTALLY AWESOME ! ! ! ! !

Compound and separate spelling

11) Faulty compound spelling:

> **Example 11: ripoff vs. rip off/rip-off**
> [. . .] universe of camelot that is a disney ripoff [. . .]

12) Faulty separate spelling:

> **Example 12: story teller vs. storyteller**
> JJ Knight is an awesome story teller [. . .]

Acronyms and abbreviations

13) Usage of acronyms:

> **Example 13: OMG**
> OMG , someone finally figured it out !

14) Abbreviations without punctuation marks:

> **Example 14: Mr Rochester vs. Mr. Rochester**
> [. . .] performance as Mr Rochester is superb [. . .]

15) Unusual abbreviations:

> **Example 15: def vs. definitely**
> [. . .] and that track is def tight

Diatopic variations and foreign languages

16) British English vs. American English:

> **Example 16: favourite (BE) vs. favorite (AE)**
> [. . .] this version is undoubtedly my favourite .

17) Foreign words:

> **Example 17: cloisonne (French: cloisonné vs. Cloisonné)**
> I 've taken a few cloisonne classes [. . .]

Stylistic features

18) Unusual discourse particles/interjections:

> **Example 18: hah**
> What author or editor would let that go to print?"Ah - hum " , " ah - huh " and " a - hah " [. . .]

19) Colloquial expressions/slang:

> **Example 19: thingamajig**
> [. . .] percentage at the bottom thingamajig

20) Alternative spelling:

> **Example 20: frikkin vs. freaking**
> [. . .] , its a frikkin tv show .

21) Neologisms:

> **Example 21: cartoonish**
> But , Lily feels cartoonish to me , [. . .]

Syntax

22) Verb in first position (declarative sentence), missing noun or pronoun:

> **Example 22: [I] Will**
> Will definitely read more of her books .

Proper nouns

23) Proper nouns:

> **Example 23: Mrs. Kuklinski**
> Winona Ryder also did a good job as Mrs Kuklinski .

24) Proper nouns vs. determinative compounds:

> **Example 24: superman vs. Superman**
> [. . .] gunshot wounds ca n't even stop our superman .

Additional features

25) Similar expressions:

> **Example 25: straightforward vs. straight forward**
> It was clear and straightforward [. . .]

26) Combination of nouns and digits:

> **Example 26: 7-year**
> [. . .] with my 7-year - old .

27) Systematic repetition of a mistake:

> **Example 27: albumns vs. albums**
> [. . .] on the live albumns [. . .]

Examples of combined linguistic features

28) Accumulated special characters:

> **Example 28: A++++++**
> Mom , you get an A++++++ from me , and I am

29) Mispelled proper nouns:

> **Example 29: Speilberg**
> Steven Speilberg 's first movie as director [. . .]

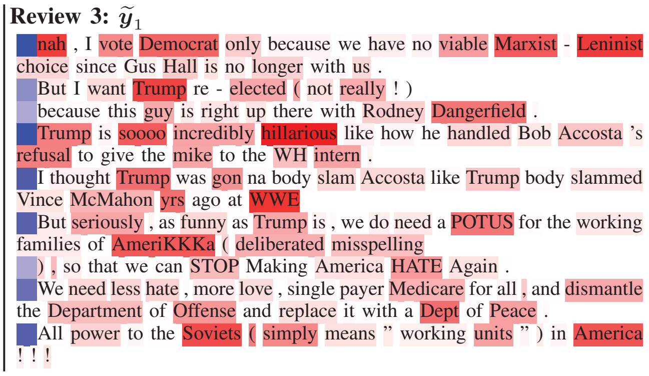
The analysis of linguistic features for the purpose of author verification is a complex research field, since there are many levels of description, such as syntax, morphology, lexicology, semantics, pragmatics, spelling etc. Concepts from corpus linguistics and variational linguistics are also significant for a comprehensive analysis. Our investigation, however, is limited by the algorithms estimation of the significance of terms and other parts of the texts. In applying an in-depth analysis of, e.g., syntactic features, whole sentences should be taken into account. When the algorithms flags an entire sentence, in turn, it is impossible to automatically determine which feature caused that decision. In addition to focusing on (topic-related) nouns, adjectives and verbs, which is also visible in Fig. 7, the algorithm favors some parts of the text over others; i.e., the last sentence of a review is frequently marked by the system.

F. Differential Example and Adversarial Example

Lastly, we want to illustrate a conceptual feature of the proposed ADHOMINEM system by returning to our example from Fig. 1. The ADHOMINEM system predicts, correctly, that both reviews from Fig. 1 were written by the same author. However, the decision is wrought with low reliability because $\tau_s = 1 < d(\boldsymbol{y}_1, \boldsymbol{y}_2) = 1.66 < \tau_d = 3$. The resulting distance of 1.66 is still below the threshold of $\frac{\tau_s + \tau_d}{2} = 2$, which leads to the correct decision. The reason for the low reliability can be found in the lack of overlap in the token set observed in each of the reviews. The overlap is only 15.7%.

An interesting differential example is constructed if we substitute the words *"cuz"* and *"w/"* with *"because"* and *"with"* in the third sentence of the first review. The updated attention-based heatmap, as produced by ADHOMINEM, can be seen in Fig. 8. As expected, our correction of the misspellings leads to a reduced attention for both words as well as for the entire sentence, which supports our claim that the chosen attentions are able to pick up on the relevance of individual linguistic features. The scoring slightly changes from $d(\boldsymbol{y}_1, \boldsymbol{y}_2) = 1.66$ to $d(\widetilde{\boldsymbol{y}}_1, \boldsymbol{y}_2) = 1.68$, so that the differential example would still be categorized correctly.

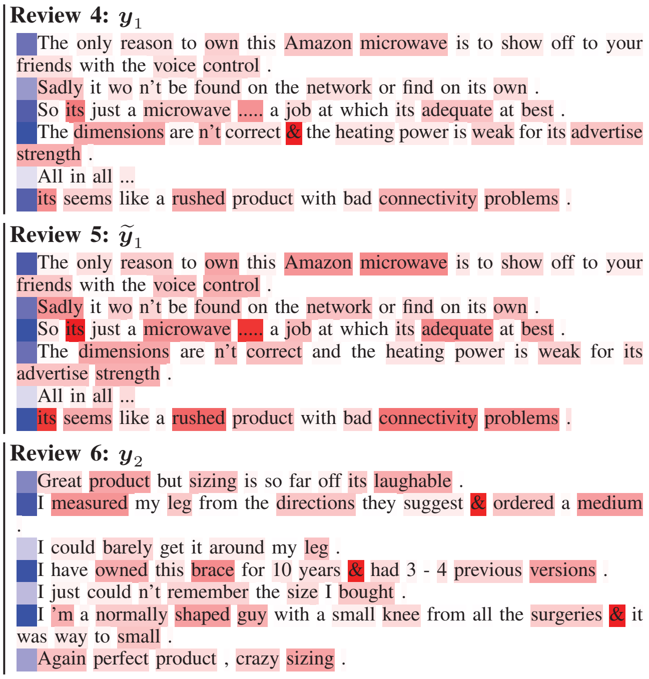
Next, we compare reviews 4 and 6 in Fig. 9. Again, both reviews were written by the same author, with a similar overlap rate of 15.1%. Our system ADHOMINEM predicted it correctly, with a score of $d(\boldsymbol{y}_1, \boldsymbol{y}_2) = 1.60$. It is observable that the author likes to use the character *"&"* instead of writing the word *"and"*. We now create an adversarial example by replacing the & character in example 4 by the word *"and"*. The updated attention-based heatmap can be seen in example 5 in Fig. 9. Again, ADHOMINEM works as expected and the word *"and"* is no longer marked as significant. As a result, however, the scoring increases to $d(\widetilde{\boldsymbol{y}}_1, \boldsymbol{y}_2) = 2.30$ which leads to a misclassification.



Review 3: $\widetilde{y}_1$

nah , I vote Democrat only because we have no viable Marxist - Leninist choice since Gus Hall is no longer with us .
But I want Trump re - elected (not really !)
because this guy is right up there with Rodney Dangerfield .
Trump is soooo incredibly hillarious like how he handled Bob Accosta 's refusal to give the mike to the WH intern .
I thought Trump was gon na body slam Accosta like Trump body slammed Vince McMahon yrs ago at WWE
But seriously , as funny as Trump is , we do need a POTUS for the working families of AmeriKKKa (deliberated misspelling
) , so that we can STOP Making America HATE Again .
We need less hate , more love , single payer Medicare for all , and dismantle the Department of Offense and replace it with a Dept of Peace .
All power to the Soviets (simply means " working units ") in America ! ! !

Fig. 8: Differential example: scoring slightly changes from $d(\boldsymbol{y}_1, \boldsymbol{y}_2) = 1.66$ to $d(\widetilde{\boldsymbol{y}}_1, \boldsymbol{y}_2) = 1.68$.



Review 4: y_1

The only reason to own this Amazon microwave is to show off to your friends with the voice control .
Sadly it wo n't be found on the network or find on its own .
So its just a microwave ..... a job at which its adequate at best .
The dimensions are n't correct & the heating power is weak for its advertise strength .
All in all ...
its seems like a rushed product with bad connectivity problems .

Review 5: $\widetilde{y}_1$

The only reason to own this Amazon microwave is to show off to your friends with the voice control .
Sadly it wo n't be found on the network or find on its own .
So its just a microwave ..... a job at which its adequate at best .
The dimensions are n't correct and the heating power is weak for its advertise strength .
All in all ...
its seems like a rushed product with bad connectivity problems .

Review 6: y_2

Great product but sizing is so far off its laughable .
I measured my leg from the directions they suggest & ordered a medium .
I could barely get it around my leg .
I have owned this brace for 10 years & had 3 - 4 previous versions .
I just could n't remember the size I bought .
I 'm a normally shaped guy with a small knee from all the surgeries & it was way to small .
Again perfect product , crazy sizing .

Fig. 9: Adversarial example: scoring changes from $d(\boldsymbol{y}_1, \boldsymbol{y}_2) = 1.60$ to $d(\widetilde{\boldsymbol{y}}_1, \boldsymbol{y}_2) = 2.30$.

Both examples exhibit the following properties: Although the review pairs have a very similar overlap rate, a simple manually performed error correction can lead to clear effects on the attention weights. On the one hand, if the reviews as in Fig. 9 are too short (with 59 and 63 tokens, respectively) it is difficult to gather enough information for a reliable decision. As a result, ADHOMINEM strongly relies on a single linguistic feature. This adversarial example shows that a very simple obfuscation strategy has the power to potentially impede correct verification, an insight which can be leveraged in future work to better understand the decision-making process for automatic verification results and for hardening such frameworks systematically against obfuscation strategies. On the other hand, considering Fig. 1 and 8, the verification of ADHOMINEM is still robust. In both reviews, the number of tokens (120 and 130) is sufficient to characterize the underlying writing styles.

V. CONCLUSION

We introduced a new algorithm for forensic text comparison called ADHOMINEM which is characterized by an attention-based Siamese network topology that is able to learn linguistic features such as spelling errors, non-standard lexical forms, and expressions that differ in other ways from the norm. We compiled a large-scale dataset of short product reviews and made it publicly accessible for text comparison tasks. Besides the quantitative evaluation of the performance of the algorithm we presented a qualitative linguistic analysis based on the visualization of the internal attention weights.

From a linguistic perspective, we grant that, despite our very encouraging results, further research may be needed. A limiting factor for any quantitative analysis in forensic linguistics is that, to our knowledge, no reliable mechanism exists to do the analysis by machine. This implies that all results that involve explainable features have to rely on a manual analysis by an expert and become, therefore, very difficult to do on large-scale datasets. Within the resources available to us we were therefore not able to conduct a comprehensive quantitative analysis of explainability. We, nevertheless, feel that our results were very encouraging and therefore worth sharing with researchers in the field.

From a technical perspective, our experiments highlight the distinct advantages of the proposed Siamese network topology over traditional methods. ADHOMINEM fuses a *self-configuring* feature extraction into a verification module to form a single framework. It is neither necessary, to manually define meaningful stylometric features, nor to acquire annotated data for preparatory steps like part-of-speech tagging. Hence, the proposed framework allows for an efficient, fully automated use of big datasets in forensic linguistics, while still retaining a large degree of interpretability, which is of special significance in this field of application. Our system demonstrates unequivocally that *big data* has become important and relevant in the field of forensic linguistic as well.

REFERENCES

[1] S. Ehrhardt, "Authorship attribution analysis," in *Handbook of Communication in the Legal Sphere*, J. Visconti, Ed. Berlin/Boston: de Gruyter, 2018, pp. 169–200.

[2] G. McMenamin, *Forensic Linguistics: Advances in Forensic Stylistics*. CRC Press, 2002.

[3] L. Solan and P. M. Tiersma, "Speaking of crime: The language of criminal justice," *Uni. of Chicago Press*, 2005.

[4] M. Coulthard, "Author Identification, Idiolect, and Linguistic Uniqueness," *Applied Linguistics*, vol. 25, no. 4, pp. 431–447, 2004.

[5] R. Shuy, *The Language of Fraud Cases*. Oxford University Press, 2016.

[6] A. Rocha, W. J. Scheirer, C. W. Forstall, T. Cavalcante, A. Theophilo, B. Shen, A. R. B. Carvalho, and E. Stamatatos, "Authorship attribution for social media forensics," *IEEE Trans. Inf. Forensic Secur.*, vol. 12, no. 1, pp. 5–33, 2017.

[7] P. Juola, "Authorship attribution," *Foundations and Trends in Information Retrieval*, vol. 1, no. 3, pp. 233–334, 2006.

[8] M. Koppel and J. Schler, "Authorship verification as a one-class classification problem," in *Proc. ICML*. ACM, 2004, pp. 62–69.

[9] M. Koppel and Y. Winter, "Determining if two documents are written by the same author," *Journal of the Association for Information Science and Technology*, vol. 65, no. 1, pp. 178–187, 2014.

[10] K. Luyckx and W. Daelemans, "Authorship attribution and verification with many authors and limited data," in *Proc. Coling*. Manchester, UK: ACL, 2008, pp. 513–520.

[11] H. J. Escalante, T. Solorio, and M. Montes-y Gómez, "Local Histograms of Character N-grams for Authorship Attribution," in *Proc. ACL*. AC, 2011, pp. 288–298.

[12] S. Afroz, M. Brennan, and R. Greenstadt, "Detecting hoaxes, frauds, and deception in writing style online," in *Proc. IEEE SSP*, 2012.

[13] S. Segarra, M. Eisen, and A. Ribeiro, "Authorship attribution using function words adjacency networks," in *Proc. ICASSP*, 2013, pp. 5563–5567.

[14] U. Sapkota, S. Bethard, M. M. y Gomez, and T. Solorio, "Not all character n-grams are created equal: A study in authorship attribution," in *Proc. NAACL*. ACL, 2015, pp. 93–102.

[15] O. Halvani, C. Winter, and A. Pflug, "Authorship verification for different languages, genres and topics," *Digital Investigation*, vol. 16, pp. S33–S43, 2016.

[16] N. Potha and E. Stamatatos, "Improving author verification based on topic modeling," *J. Assoc. Inf. Sci. Technol*, 2019.

[17] E. Frank, C. Chui, and I. H. Witten, "Text categorization using compression models," in *Proc. DCC*, 2000, pp. 555–.

[18] E. Stamatatos, "A survey of modern authorship attribution methods," *J. Assoc. Inf. Sci. Technol.*, vol. 60, no. 3, pp. 538–556, 2009.

[19] S. H. H. Ding, B. C. M. Fung, F. Iqbal, and W. K. Cheung, "Learning stylometric representations for authorship analysis," *IEEE Trans. on Cybern.*, vol. 49, no. 1, pp. 107–121, 2019.

[20] J. Hitschler, E. van den Berg, and I. Rehbein, "Authorship attribution with convolutional neural networks and POS-Eliding," in *Proc. StyleVar*. ACL, 2017, pp. 53–58.

[21] P. Shrestha, S. Sierra, F. Gonzalez, M. Montes, P. Rosso, and T. Solorio, "Convolutional neural networks for authorship attribution of short texts," in *Proc. EACL*. ACL, 2017, pp. 669–674.

[22] D. Bagnall, "Author identification using multi-headed recurrent neural networks." in *Proc. CLEF*, 2015.

[23] M. Litvak, "Deep dive into authorship verification of email messages with convolutional neural network," in *Proc. SIMBig*. Springer, 2018, pp. 129–136.

[24] A. Thephilo, L. A. M. Pereira, and A. Rocha, "A needle in a haystack? harnessing onomatopoeia and user-specific stylometrics for authorship attribution of micro-messages," in *Prc. ICASSP*, 2019, pp. 2692–2696.

[25] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-speech tagging for twitter: Annotation, features, and experiments," in *Proc. ACL*. ACL, 2011, pp. 42–47.

[26] R. Schwartz, O. Tsur, A. Rappoport, and M. Koppel, "Authorship attribution of micro-messages," in *Proc. EMNLP*. ACL, 2013, pp. 1880–1891.

[27] B. Boenninghoff, R. M. Nickel, S. Zeiler, and D. Kolossa, "Similarity learning for authorship verification in social media," in *Proc. ICASSP*, 2019, pp. 2457–2461.

[28] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. NAACL*, 2016, pp. 1480–1489.

[29] X. Ma and E. Hovy, "End-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF," in *Proc. ACL*. ACL, 2016, pp. 1064–1074.

[30] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015.

[31] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," in *Proc. AAAI*. AAAI Press, 2016, pp. 2786–2792.

[32] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. EMNLP*. ACL, 2014, pp. 1746–1751.

[33] 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.

[34] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, "Image-based recommendations on styles and substitutes," in *Proc. SIGIR*. ACM, 2015, pp. 43–52.

[35] L. van der Maaten and G. E. Hinton, "Visualizing high-dimensional data using t-SNE," *JMLR*, vol. 9, pp. 2579–2605, 2008.

[36] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Proc. LREC*, 2018.

[37] M. Hürlimann, B. Weck, E. van den Berg, S. Suster, and M. Nissim, "GLAD: Groningen Lightweight Authorship Detection," in *Proc. CLEF*, 2015.

[38] O. Halvani, C. Winter, and L. Graner, "Unary and binary classification approaches and their implications for authorship verification," *CoRR*, vol. abs/1901.00399, 2019.

[39] S. Jain and B. C. Wallace, "Attention is not explanation," in *Proc. NAACL*. ACL, 2019.

[40] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein, "Explainable prediction of medical codes from clinical text," in *Proc. NAACL*. New Orleans, Louisiana: ACL, 2018, pp. 1101–1111.