Representation learning of writing style

Julien Hay[1,2,3], **Bich-Liên Doan**[2,3], **Fabrice Popineau**[2,3] **and Ouassim Ait Elhara**[1]

[1]Octopeek SAS, 95880 Enghien-les-Bains, France
[2]Laboratoire de Recherche en Informatique, Paris-Saclay University, 91190 Gif-sur-Yvette, France
[3]CentraleSupélec, Paris-Saclay University, 91190 Gif-sur-Yvette, France
{julien.hay, ouassim.aitelhara}@octopeek.com
{bich-lien.doan, fabrice.popineau}@centralesupelec.fr

Abstract

In this paper, we introduce a new method of representation learning that aims to embed documents in a stylometric space. Previous studies in the field of authorship analysis focused on feature engineering techniques in order to represent document styles and to enhance model performance in specific tasks. Instead, we directly embed documents in a stylometric space by relying on a reference set of authors and the *intra-author consistency* property which is one of two components in our definition of writing style. The main intuition of this paper is that we can define a general stylometric space from a set of reference authors such that, in this space, the coordinates of different documents will be close when the documents are by the same author, and spread away when they are by different authors, even for documents by authors who are not in the set of reference authors. The method we propose allows for the clustering of documents based on stylistic clues reflecting the authorship of documents. For the empirical validation of the method, we train a deep neural network model to predict authors of a large reference dataset consisting of news and blog articles. Albeit the learning process is supervised, it does not require a dedicated labeling of the data but it relies only on the metadata of the articles which are available in huge amounts. We evaluate the model on multiple datasets, on both the authorship clustering and the authorship attribution tasks.

1 Introduction

Authorship analysis is an ensemble of methods that aims to extract useful authorship information of a text by analyzing writing style. The most commonly addressed tasks in this field are classification tasks such as authorship attribution (Stamatatos, 2017), authorship verification (Boumber et al., 2019) and authorship characterization. The authorship attribution is the process of inferring the author of documents among known authors while the authorship verification is the process of deciding whether or not a given document was written by a given author. To this end, most studies rely on feature engineering to represent the input documents in order to improve the performance of machine learning algorithms. Feature engineering consists in transforming raw input data into new input data by using domain knowledge with the expectation that the new features will be more suitable for learning an efficient model. For the two aforementioned tasks, the process consists in selecting textual characteristics of documents, then, either use a classifier to predict the author of a document based on these characteristics, or calculate similarities between document representations.

One common way to choose these document representation features is by assessing whether or not they can enhance the prediction accuracy. Sometimes these features intuitively belong to style such as function words (Goldstein-Stewart et al., 2009; Menon and Choi, 2011), sometimes they just correspond to common NLP features such as distributional representations of documents (Chen et al., 2017; Gupta et al., 2019; Bagnall, 2015). Few studies attempt to directly produce unsupervised representations of style in order to project unseen documents in a low dimensional stylometric space (Ding et al., 2019; Jasper et al., 2018; Boumber et al., 2019). Feature engineering is designed by humans based on heuristics, a labor-intensive process. However, recent studies show that automatically learning a representation of raw data is useful for many reasons described in Bengio et al. (2013). Learned representations are suitable in classification and clustering tasks since these representations manage to select discriminating information from raw data and represent them in a low dimensional vector space (Arora and Risteski, 2017). Moreover,

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 232–243
Online, Nov 19, 2020.

this discriminating information is not necessarily captured by humans through heuristics.

Documents belonging to same authors are generally consistent in their writing style (Karlgren, 2004) even for authors covering a large range of topics (Patchala and Bhatnagar, 2018). The method we propose relies on this observation. In this paper, we validate the *style-generalization* assumption which is based on two propositions. First, it states that we can represent unseen documents of unseen authors by generalizing stylometric features from a set of known authors and known documents. Second, documents that belong to the same author tend to have similar representations in the stylometric space (given a standard similarity function). In order to validate this assumption, we exploit recent advances in text and sentence representation with deep neural network (DNN) architectures, especially transformers (Devlin et al., 2019; Yang et al., 2019). We propose a transformer-based DNN model fine-tuned on the authorship attribution task using a large dataset of documents whose authors are known (the reference set). Then we assess the model in its ability to capture the similarity between documents of the same authors in a clustering experiment.

In this paper, we first give an overview of related work on authorship analysis in Section 2. In Section 3, we propose a definition of the style and the motivations behind our representation learning method. In Section 4, we explain our method and introduce the *style-generalization* assumption. Section 5 presents the experimentation on the main task which is the authorship clustering and on a second, the authorship attribution task. Finally, in Section 6, we deepen our analysis by studying the second property of our definition of style.

2 Related work

Neal et al. (2017) and Stamatatos (2009) gave an overview of features used for authorship analysis. Categories of features for stylometry are lexical (e.g. sentences length, vocabulary richness), syntactic (e.g. punctuation, *Part-of-Speech* tags), semantic (e.g. synonyms, semantic dependencies), structural (e.g. average paragraph length, presence of quotes) and application-specific (presence of words in a specific lexicon). Authors also consider additional features which are hard to classify such as topic modeling based features and readability metrics. For instance, Bayesian methods such as LDA was shown to be efficient in the e-mails and blog content authorship attribution (Seroussi et al., 2014) and the research paper authorship attribution (Rosen-Zvi et al., 2004). Most experiments in authorship analysis involve training machine learning models that take as input different features of these categories (Houvardas and Stamatatos, 2006; Yang et al., 2018; Tausczik and Pennebaker, 2010).

Recent studies tackle issues of the feature engineering process for authorship analysis by exploiting raw text samples using deep neural networks. Chen et al. (2017) proposed a gated recurrent unit (GRU) DNN trained on article and sentences for the authorship verification task. Other studies demonstrated the relevance of recurrent neural networks such as GRU and long short-term memory (LSTM) DNNs on the authorship attribution task (Gupta et al., 2019; Bagnall, 2015). More recently, Boumber et al. (2018) proposed a convolutional neural network-based model for the multi-label authorship attribution of scientific publications.

Besides the use of DNN for classification, very few studies attempt to automatically embed stylometric features from a corpus of documents. Qian et al. (2015) are the first to rely on an external dataset of known authors to pretrain a general model which can compute stylometric similaries between documents. They proposed training a support vector machine-based (SVM) model that was re-used for a test dataset containing unseen authors. As far as we know, this study is the first attempt at generalizing a stylometric similarity space while still relying on handcrafted stylometric features. Ding et al. (2019) proposed a model that jointly learns topical and lexical distributional representation of documents in an unsupervised manner to help authorship analysis. Jasper et al. (2018) proposed a model that embeds the writing style of English novels. The model is composed of *fastText* word embeddings and a stacked LSTM. The authors demonstrated that their model performed well on a subset of the PAN14 dataset for the authorship verification task. In order to verify authorship of social networks posts, they designed their experiments so that model inputs were short text samples. Boumber et al. (2019) recently used a recurrent neural network-based architecture and adversarial learning to tackle the authorship verification task. The model was trained to embed pairs of documents and was assessed on authorship verification task in transfer learning settings, when

authors of the test set do not exist in the train set.

Instead, we propose a representation learning method that aims to embed documents in a stylometric space relative to a large dataset of well-known authors. We use a modification of a pre-trained BERT model (Sanh et al., 2019) in a classification task as proposed in several recent works (Sun et al., 2019; Reimers and Gurevych, 2019). To the best of our knowledge, this paper is the first to propose an authorship attribution-training driven model generalizing stylometric features so that documents of unseen authors can be clustered without fine-tuning. Additionally, our method relies on a large dedicated dataset that can be extended. We trained our models on a large amount of data by using news and blog articles benefiting from the wide availability of such data on the web.

3 Motivation

An author can adopt several styles, and one of them can be similar to the writing style of another author. Karlgren (2004) defined the style as "a consistent and distinguishable tendency to make [some of these] linguistic choices". Moreover, Karlgren (2004) explained that "texts are much more than what they are about". Any textual characteristic that is not semantic or topical belongs to stylistic choices of the author. Different expressions can have a common meaning, and can refer to the same objects and the same events, but still be made up of different words and different syntax, corresponding to the author's willingness to let a context, an orientation, sometimes an emotion be shown through (Argamon et al., 2005).

News articles showed to have specific writing style by using, for instance, date as adverb in "The governor Thursday announced..." or anthropomorphization in "The 1990s saw an increase in crime...". It is called journalese (Dickson and Skole, 2012). Headlines of newspapers also have their own style, called headlinese, such as articles drop (Weir, 2009). Today's trend towards clickbait also has an influence on headlines (Chakraborty et al., 2016). Finally, let's mention style guides of newspapers that not only influence typography (e.g. paragraph structure, quotation, italic) but also the usage of the language (Cameron, 1996). For instance, a style guide can more or less encourage the use of the active voice against the passive voice. Some of the newspapers can have their own style guide that their writers follow, allowing the text to be consistent for the reader while any variation having no purpose could be distracting (Hicks, 2002).

Style appears more or less pronounced depending on the text passages, it is difficult to define it precisely and, given a document, to find a set of words (or sequence of words) that will strictly define the style of its author. The text is the combination of a shape – its style – and a content which are intertwinned thanks to the choice of specific words. Words or sequence of words in the text can rarely be denoted as belonging specifically to the style or to the content. This is why extracting style features is hard. From documents of a reference corpus, we aim to extract latent structures falling within the scope of writing style. We argue that these latent structures can be identified by DNNs, typically RNN models with attention layers which will focus on style-related terms. From a linguistic point of view these latent structures map to lexical, syntactic or structural fragment of sentences or paragraphs. Intuitively, when extracting a style representation of a document, we seek to focus on latent structures that will satisfy these two properties :

Intra-author consistency the property of being consistent in documents belonging to the same author.

Semantic undistinguishness the property of carrying very little information on what makes the document semantically (e.g. topics, named entities) distinguishable in the corpus.

Thus, this definition, inspired by Karlgren (2004); Holmes (1998), means that the style of a document is represented by linguistic structures which are consistent for individual authors (allowing their identification) but more likely semantically poor regarding the content of the document (e.g. topic, named entities). Indeed, what the document is about is a constraint that imposes on the author to use a specific vocabulary. The terms that belong to this specific vocabulary have a strong semantic value with respect to the theme of the document, and on the contrary, are less likely to convey the author's style. The representation learning method is based on identifying consistent latent structures following the *intra-author consistency* property. Next to that, the *semantic undistinguishness* is a property which can be verified by studying attention weights of a trained DNN models.

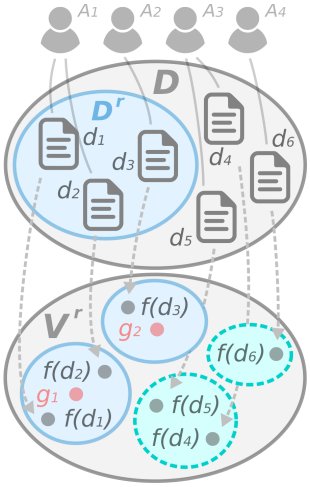


Figure 1: The *style-generalization* assumption

4 Method

Let's denote $D = \{d_1, ..., d_n\}$ a set of documents and $A = \{a_1, ..., a_m\}$ a set of authors so that each document belongs to one and only one author and each author wrote at least one document.

Let's denote $R\text{-}set = (D^r, A^r)$ the reference set with $D^r \subset D$, $A^r \subset A$, $|D^r| = n^r$, $|A^r| = m^r$ and A^r is the set of all the authors of the documents in D^r. Sizes n^r and m^r are typically large and $n^r \geq m^r$.

Let's denote $V^r \in [0,1]^{m^r}$ a vector space so that $\forall v \in V^r, \sum_{i=1}^{m^r} v_i = 1$. Thus, softmax vectors for A^r belong to V^r. We denote $G^r = \{g_1^r, ..., g_{m^r}^r\}$ the set of one-hot vectors that correspond to the ground-truth vectors of each author in A^r. Thus, vectors from G^r also belong to V^r. Each document in D^r is associated with one and only one vector from G^r.

Let $f^r : \mathbb{R}^l \times \mathbb{R}^w \rightarrow V^r$ a function that projects input documents, represented by word vectors, from D^r into V^r with l the size of the documents (truncated or padded) and w the dimension of the word vectors. The purpose of f^r is to project each document $d_i^r, i \in \{1, ..., n^r\}$ closer to its corresponding vector in G^r than to any other ground-truth vector.

The main assumption of this paper, the *style-generalization* assumption, is that two representations defined by f^r of two unseen documents are more likely to be similar if they belong to the same author than if they do not. Intuitively, we assume that any unseen document belonging to an unknown author is similar, in terms of style, to documents belonging to a subset of known authors and that another document of the same unknown author is likely to be similar, in terms of style, to documents belonging to this same subset of known authors. Figure 1 illustrates the *style-generalization* assumption. Blue sets in V^r correspond to representations of documents from D^r. f^r allows to project documents from D^r to V^r such that the representation of a document is close to its corresponding ground truth vector (in red color). Green sets correspond to unseen document representations in V^r. The assumption states that close representations (green sets) likely belong to the same author.

More formally, let's denote $U\text{-}set = (D^u, A^u)$ a set of unseen documents and authors with $D^u \subset D$, $A^u \subset A$, $|D^u| = n^u$, $|A^u| = m^u$ and A^u is the set of all the authors of the documents in D^u. $A^r \cap A^u = \emptyset$ and $D^r \cap D^u = \emptyset$. The *style-generalization* assumption states that the projection of documents from D^u (the $U\text{-}set$) into V^r using f^r allows to compute similarities such that similar documents from D_u are likely written by the same author. Thus, learning f^r on a reference set allows authorship clustering in the general stylometric space that it defines.

This method, which can be called the *style-generalization* learning method, aims to calculate stylometric similarities of documents without considering the author of the document, i.e. without fine-tuning. In Section 5, we propose to use a DNN model as the function f^r. We train the model on the reference set and generate vector representations of unseen documents by using intermediate weights representation in the DNN.

5 Experimentation

5.1 Dataset

Since it is important to have heterogeneous documents in the dataset to generalize representations, we relied on a large amount of English news and blog articles. We merged all documents from The Blog Authorship Corpus (Schler et al., 2006), ICWSM datasets (Burton et al., 2009, 2011) and news collected for this study[1].

All documents have at least one domain name such as *nytimes.com*. Authors were extracted from the HTML content. In cases when an author is found, we consider the label of the document to be the concatenation of the domain and the author, or else the label is the domain alone. Domains correspond, in most cases, to online newspapers or blogs.

[1]Datasets, code and pretrained models are available at https://github.com/hayj/DeepStyle

We made the final dataset, named *NewsID* (*News Source Identification*), by only keeping documents longer than 20 words. We also removed documents having a class (the author) that is sub-represented, i.e. a class with less than 200 documents.

Finally, we randomly generated a $R\text{-}set$ gathering $\sim$3.3 millions of documents and 1 200 different classes. We set a limit of 3 0000 documents per class. The average number of documents per class is 3 000. The $R\text{-}set$ is large which is a requirement for our method. It is composed of reference authors with balanced numbers of documents to avoid having majority classes and majority "reference styles". We also randomly generated $\sim$500 *U-sets* having 50 classes and 50 documents per class. *U-sets* does not contain any authors or documents of the $R\text{-}set$. Each *U-sets* is specific to an online newspaper or a blog (e.g. *blogger.com*, *livejournal.com*, *washingtonpost.com*, *breitbart.com*, *cnn.com*, *theguardian.com* and *nytimes.com*) and gather documents of different authors from the given websites. For example, the *U-set nytimes.com* has 2 500 news articles of *nytimes.com*. The articles were written by authors of this online newspaper. Each author wrote 50 articles in the *U-set*. The same goes with other online newspaper and blogs.

In the *NewsID* dataset, we concatenated the authors of the articles with the domain name of online newspapers and blogs. After studying the labels of our dataset, we noticed that a large majority (approximately 99%) of the authors common to several journals are actually namesakes because they use short pseudonyms (*Alex*, *Erik*, *Lucy*, etc.). Thus, we have not differentiated the rare cases where an author appears to have written for several newspapers (e.g. *Andrew Restuccia* for *politico.com* and *thehill.com*) by manual labeling. Authors are consistent in their writings, and as we have seen, so are online newspapers, especially because their writers have to follow a style guide. But even in cases where it is difficult to distinguish the articles of an author who has written for two different newspapers, we argue that the existence of two labels for the same author will have little impact on the learning of our models and the stylometric representation of the documents. Indeed, if the model fails to capture the differences between the documents of this author, the consequence will be that the internal representations of the deep neural network will be close for the documents of one or the other of the two labels it has to predict. The similarity of the internal representations of the deep neural network for these two reference authors will not necessarily imply the addition of noise in the projection of unseen documents from the *U-sets*.

5.2 Models for learning f^r

In order to validate the *style-generalization* assumption, we need to learn the function f^r. To fulfill this objective, we propose to train two DNNs on the $R\text{-}set$ following the authorship attribution task, a classification task. Data are documents of the *R-set* and labels are those described in Section 5.1. By relying on the *intra-author consistency* property, we train the models to capture consistent lexical clues of each author. Then we rely on the trained model to embed documents of the $U\text{-}set$ by selecting weights in an intermediate layer.

First we implemented the *SNA* model (*Stylometric Neural Attention*) which is a bi-directionnal LSTM with attentions mainly based on the architecture proposed by Zhou et al. (2016) with two fully connected layers of 500 units and of softmax layer of 1 200 units. The loss function is the multi-class log loss. We set dropouts of each layer to 0.2. Inputs of the DNN are the *GloVe 840B* word vectors of firsts 1 200 words (Pennington et al., 2014) of a document. Documents are padded or truncated so that each has a size of 1 200 words.

Second, we implemented a model based on the BERT transformer architecture (Devlin et al., 2019) which is a bidirectional attention model with the use of masked words and next sentence prediction for the unsupervised pre-training phase. We use a variant of BERT, called DistilBERT, which uses the knowledge distillation principle reducing the size of the final DNN model (Sanh et al., 2019). We used DistilBERT because its training is less time-consuming. The DistilBERT base model was trained on an English corpus of books and Wikipedia pages. The *DBert-ft* model corresponds to an authorship attribution fine-tuning of the DistilBERT base model (uncased) on the *R-set*. We used a linear layer on top of the base model and a multi-class log loss as the loss function. The size of the last layer is 1 200 and corresponds to the number of classes in the *R-set*. We let all layers trainable. Dropouts are set to 0.1.

When training the *DBert-ft* model, input documents are split into multiple parts of 512 *word-pieces*. Each part has the same label. This allows to increase the number of samples in the *R-set*. Inputs of the DNN take the form of 512 indexes

from a common *wordpieces* vocabulary. In the next sections of this paper, when generating the representation of a single document, we compute the mean vector of all its parts. By doing so, we capture more information that can be used in the representation of the authorship of the document than if we only considered the first or last part of the document. Vector representations are outputs of the last layer before the classification layer on top of the model.

Intermediate layer choice for representation of documents in both models as well as the mean of document parts have been experimentally validated on a validation *U-set*. Both models are implemented with *TensorFlow* (Abadi et al., 2015). For both models, the learning time was about one week on a *NVIDIA TITAN V* GPU (12GB memory).

5.3 Baselines

In order to compare representations of our models in the authorship clustering experiment, we use several baselines. We generate random vectors of different dimensions and variances. ***Stylo*** corresponds to stylometric handcrafted features commonly used in authorship analysis such as readability scores, vocabulary richness, sentences count. ***TFIDF*** corresponds to TFIDF weights of documents reduced to 100 dimensions using SVD. ***LDA*** (Blei et al., 2003) corresponds to topics vectors (100 topics) of documents. ***Doc2Vec*** corresponds to vectors representations of a *Doc2Vec* (Le and Mikolov, 2014) model trained in an unsupervised manner on documents of the *NewsID* R*-set*. ***USent*** corresponds to vectors of the *Universal Sentence Encoder* model trained on English Wikipedia and news data (Cer et al., 2018). ***InferSent*** corresponds to vectors of an InferSent model trained on natural language inference data (Conneau et al., 2017). ***BERT*** corresponds to vectors of BERT trained on large English books and wikipedia articles dataset (Devlin et al., 2019). We used the large and uncased version of BERT. ***DBert*** corresponds to vectors of the non fine-tuned DistilBERT uncased model (Sanh et al., 2019). For models producing sentence representations (*USent*, *InferSent*, *BERT*), inputs are the mean of sentences representations.

5.4 Metrics

In order to evaluate each model, we first generated vector representations $V = \{v_1, ..., v_p\}$ of all $d_p^u \in D^u$ (the $U\text{-}set$) using a given model such that $\forall v \in V, |v| = k, k > 0$. Then, we assessed the ability of these representations to cluster the documents well according to a similarity measure and ground truth labels (internal clustering evaluation). We use labels $L = \{L_1, ..., L_p\}$ such that $L_i = L_j$ if and only if d_i^u has the same author as d_j^u. We rely on a standard clustering metric: the *Davies-Bouldin index* (*DavB*) (Davies and Bouldin, 1979). *DavB* takes V and L and returns a clustering quality score greater than 0. It is defined as the average similarity between each cluster and its most similar one. The lower the *DavB* is, the better is the quality of clusters.

In this experiment, we also want to assess, on average, how well documents are ranked in relation to each other, given their vector representation and a similarity measure. Thus we introduce a new clustering metric, called *SimRank*, based on a commonly used metric assessing the ranking quality: the *nDCG* (Järvelin and Kekäläinen, 2002).

$$\text{SimRank}(REL) = \frac{\sum_{p=1}^{|REL|} \text{nDCG}'(REL_p)}{|REL|} \quad (1)$$

Equation 1 gives the *SimRank* with REL a set of ranking vectors. A ranking vector (or graded relevance vector) rel is a vector such that $|rel| = k$ with k the number of documents in the *U-set* and $rel_i \in \{0, 1\}$. rel_i indicates whether the corresponding document in the ordered set of documents is relevant ($rel_i = 1$) or not ($rel_i = 0$). Documents are ordered by the cosine similarity between their vector representations and the vector representation of a target document. In this experiment, a document is relevant if it belongs to the same author as the target document. Thus, in our case, each rel vector corresponds to a ranking vector of each document in the *U-set* ordered by similarity with a target document in the *U-set*. REL is a square matrix corresponding to the all ranking vectors given each of documents in the *U-set* as the target document.

$$\text{DCG}(rel) = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)} \quad (2)$$

$nDCG$ is the DCG given in equation 2 normalized between 0 and the DCG of the ideal ranking (namely the $iDCG$). In Eq. 1, we use $nDCG'$ which is the DCG normalized between the DCG of the worst ranking and the DCG of the ideal ranking. We normalized $nDCG$ using the worst ranking because, in our case, we always rank every

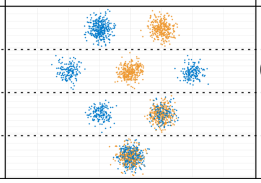
Id	Clusters	SimRank (±2σ)	DavB (±2σ)
1		1.0 (±0)	0.25 (±0.01)
2		0.69 (±0.01)	170 (±393)
3		0.66 (±0)	1.27 (±0.05)
4		0.53 (±0)	38 (±109)

Figure 2: Average *SimRank* and *DavB* scores of 1 000 randomly generated sets of 600 samples in four different cluster configurations

document according to a target document, while $nDCG$ is usually computed on a subset of documents that are returned by a search engine. Thus, the minimal value if we had used the original $nDCG$ would not be zero.

Figure 2 shows the advantages of *SimRank* compared to *DavB*. It shows the average *SimRank* and *DavB* scores of 1 000 randomly generated sets of 600 samples in four different cluster configurations. For the same cluster configuration, *SimRank* has less variance than *DavB*. In addition, some clusters such as those in the second configuration will be considered of lower quality than those in the last two configurations by the *DavB* metric.

However, as we discussed in section 3, an author can adopt several styles. For example, an author writing on political topics may sometimes write articles in a factual and descriptive style, and occasionally write in a completely different style: e.g. humorous, satirical. Thus, the articles of this author will be divided into two distinct clusters in a stylometric space as in the second configuration in the Figure 2. By using *SimRank*, we do not want to penalize these cases. Taking into account the order of the examples and not the distances, as well as the fact that the weights attributed to the samples decrease with the use of *nDCG*, allows us to minimize the effects of this multi-partitioning.

The *DavB* and the *SimRank* metric will give an insight of models performance regarding external tasks such as classification tasks. After the evaluation of models using both clustering metrics, we will evaluate models on the authorship attribution task which consist in predicting right author labels in *U-sets* (classification task). We will use the accuracy metric.

5.5 Results

Table 1 shows the performance of all models and baselines for the *SimRank* and *DavB* metrics. Scores are the mean on 22 *U-sets*. *DBert-ft* and *SNA* score higher on the *SimRank* metric following by *TFIDF* and *Doc2Vec*. *DBert-ft*, *SNA* and *TFIDF* also perform well for the *DavB* metric with close scores. In this experiment, all documents and authors of test sets (*U-sets*) are unknown for all evaluated models.

Model	SimRk	DavB
Random	0.185	14.81
TFIDF	0.455	**4.683**
LDA	0.309	8.353
Stylo	0.276	65.71
Doc2Vec	0.430	6.194
USent	0.416	5.328
InferSent	0.374	5.625
BERT	0.378	5.469
SNA	0.463	4.785
DBert	0.339	7.058
DBert-ft	**0.474**	4.777

Table 1: Authorship clustering on 22 $U\text{-}sets$

All parameters, such as dimensions, number of topics and window size of *Doc2Vec*, were grid-searched on a validation *U-set*. In this experiment, we also generated *Sent2Vec* vectors (Pagliardini et al., 2018) but we didn't add it in results since the others sentences representation models score higher. The same goes for the non-negative matrix factorization on TFIDF weighting, which scores lower compared to *LDA*. Regarding the *SNA* model, we implemented a version without an attention layer and another with an unidirectional LSTM. We obtained lower scores for each of them. The difference in scores of *DBert* and *DBert-ft* indicates that the proposed *style-generalization* learning method allows to train a model generating better authorship clusters.

In addition to the authorship clustering, we propose to compare all these models on the authorship attribution task. Table 2 shows the mean accuracy of all combinations of models on the same 22 *U-sets*. The diagonal gives scores of models alone. In order to evaluate each of these combinations on a given *U-set*, we trained a linear SVM classifier model on 80% of the *U-set* with the concatenation of vector representations as input data. The score corresponds to the accuracy of predicting the right author label on the 20% remaining data. The model choice and its hyperparameters were grid-searched on another validation *U-set*.

Results show that the *DBert-ft* model obtains the

Model	*TFIDF*	*LDA*	*Stylo*	*Doc2Vec*	*USent*	*InferSent*	*BERT*	*SNA*	*DBert*	*DBert-ft*
TFIDF	0.514	0.525	0.096	0.475	0.599	0.629	0.553	0.581	0.547	0.598
LDA		0.163	0.098	0.477	0.518	0.590	0.541	0.555	0.541	0.600
Stylo			0.098	0.108	0.103	0.097	0.102	0.101	0.097	0.191
Doc2Vec				0.472	0.474	0.491	0.526	0.543	0.519	**0.641**
USent					0.499	0.612	0.538	0.579	0.550	0.598
InferSent						0.594	0.560	0.598	0.578	0.604
BERT							0.536	0.621	0.571	0.616
SNA								0.552	0.598	0.614
DBert									0.522	0.610
DBert-ft										**0.597**

Table 2: Mean accuracy (22 *U-sets*) of models combinations for the authorship attribution task

best accuracy when used alone. Its combination with *Doc2Vec* vectors obtain the highest scores, showing that these two representations are complementary and are able to capture different clues for the classification of authors. Note that *InferSent*, despite its low scores on the authorship clustering, obtains scores close to those of *DBert-ft* when used alone. Its combination with *TFIDF* vectors also obtains scores near the best combination. We tested the same combinations of models on the authorship clustering task with the *SimRank* metric and obtained the same results, i.e. *DBert-ft* got the highest scores when combined with the other models, and the best combination was *DBert-ft* with *Doc2Vec*. All these results validate the *style-generalization* assumption and prove the benefit of the method.

6 In-depth analysis

In this section, we intend to assess how well the trained models can, by using the *style-generalization* learning method, focus on terms exposing the second property of our definition of writing style: the *semantic undistinguishness*. For this purpose, we propose to analyze attention weights of the *SNA* model. In this experiment, our main concern was not the performance in the authorship attribution task but on the use of the attention layer trained with the method described in Section 4. Thus, we used *SNA* since its training phase is less time-consuming.

The *semantic undistinguishness* suggests that style-related linguistic structures tend to carry little information on content, topics, entities, etc. These style-related structures are often referred to as function words which are frequent in a corpus (Kestemont, 2014; Argamon et al., 2007). On the other hand, terms with a high semantic value that will identify, for instance, a topic, are those allowing the document to be distinguishable in a corpus. The TFIDF weighting is a well established method to estimate how important a word is to a document in a corpus. Thus, in order to quantitatively assess the *undistinguishness* of the *SNA* model, we propose a measure based on the TFIDF weighting. The *TFIDF focus* measure allows to compute how well attentions of the model focus on words having lower TFIDF weights:

$$\text{TFIDFFocus}(A, T) = \frac{\sum_{i=1}^{d} \sum_{j=1}^{w} A_{ij}.T_{ij}}{d} \quad (3)$$

A is the attention matrix of size $w \times d$. w is the number of words in a document that we set to 1200 and d is the number of documents. Each line of the matrix corresponds to the attention weights in the *SNA* model for a document in a given *U-set*. An attention vector of a single document is normalized so that the weights sum to 1. The same goes with the normalized TFIDF matrix T of size $w \times d$. Thus, we defined the *TFIDF focus* as the mean of the attention weight times the TFIDF weight of each word. This measure is high when high values of TFIDF are in line with high values of attention and low when these high values of TFIDF are in line with low values of attention.

Given a pretrained *SNA* model and a *U-set*, we generate the matrix A using the model, the matrix T using the TFIDF weighting on the target *U-set* and, finally, the *TFIDF focus* score. Table 3 reports *TFIDF focus* scores on five *U-sets* composed of news articles and five *U-sets* composed of blog arti-

U-set type	*SNA* trained on	*U-set 1*	*U-set 2*	*U-set 3*	*U-set 4*	*U-set 5*	*Mean*
News	***Target U-sets***	0.642	0.650	0.606	0.601	0.661	0.632
	Other U-set	0.611	0.591	0.576	0.559	0.623	0.592
	R-set	**0.497**	**0.479**	**0.477**	**0.459**	**0.507**	**0.483**
Blog	***Target U-sets***	0.668	0.722	0.734	0.645	0.702	0.694
	Other U-set	0.637	0.670	0.670	0.606	0.648	0.646
	R-set	**0.547**	**0.579**	**0.575**	**0.526**	**0.560**	**0.557**

Table 3: *TFIDF focus* of *SNA* models on 5 news *U-sets* and 5 blog *U-sets*.

cles. For news articles, the third line shows *TFIDF focus* scores of the original *SNA* model trained in Section 5 following the *style-generalization* learning method. The first line shows *TFIDF focus* scores computed by a *SNA* model trained on the target *U-set*. Thus, these models learn to focus on words specific to authors in the target *U-set* to perform well in the authorship attribution. Scores show that these words have higher TFIDF weights. The second line shows *TFIDF focus* scores computed by a *SNA* model trained on an external *U-set* that we randomly chose. The external *U-set* acts as a short *R-set* with fewer documents and authors. As we can see, the use of a short *R-set* is not sufficient for the model to focus on words with lower TFIDF weights. The same goes for the three last lines but for blog articles.

In this experiment, we quantitatively showed that the original *SNA* model focuses its attention on function words having lower TFIDF weights (10% less on average), thus it is more able to capture stylometric features related to specific words exposing the *semantic undistinguishness* property than other models trained on smaller *U-sets*. The method proposed in Section 4 as well as the use of a large reference corpus have an impact on the results.

Note that, in the clustering experiment in Section 5, good performances of standard baselines such as *TFIDF* can be explained by the fact that documents of same authors have a topic bias (they share same semantic/topic words) because an author generally write on a few topics. This bias helps representations of vocabulary-based models to be close. However, such features fail to identify authors in cross-domain scenarios (Stamatatos, 2018), while our model focuses less on topic- and semantic-related words but achieves comparable performance (even better for *SimRank* and the authorship attribution task).

7 Conclusion

In this paper, we proposed a new method for the representation learning of writing style. We have shown that it is possible to generalize the writing style on the basis of a set of reference authors. The method follows a property of the style that we call *intra-author consistency*. We sought to validate two underlying propositions of the *style-generalization* assumption. First, we can represent unseen documents of unseen authors by using a model generalizing stylometric features from a set of known authors and known documents. Second, if two unseen documents have close representations using this model, they are likely to belong to the same author. Results show that the DNN model that was trained following our method succeeded in the authorship clustering of unseen documents belonging to unseen authors. It also performs well on the authorship attribution task. Moreover, we showed that a model trained with the *style-generalization* learning method is more able to capture stylometric structures exposing the *semantic undistinguishness* property.

From a practical point of view, our method does not require a tedious labeling effort but relies only on the metadata of a large dataset of articles. We believe that this work provides new perspectives in the field of authorship analysis by proposing a definition of writing style based on distributional properties, as well as a new method aiming to learn stylometric representations. In further studies, we intend to exploit style features in external tasks such as news recommendation. We already have promising results showing that style representations of news articles allow to diversify recommendation lists and to recommend "novel" news articles without loosing prediction accuracy.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James W. Pennebaker. 2005. Lexical predictors of personality type. In *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*.

Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822.

Sanjeev Arora and Andrej Risteski. 2017. Provable benefits of representation learning. *CoRR*, abs/1706.04601.

Douglas Bagnall. 2015. Author identification using multi-headed recurrent neural networks. *CoRR*, abs/1506.04891.

Y. Bengio, A. Courville, and P. Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Dainis Boumber, Yifan Zhang, Marjan Hosseinia, Arjun Mukherjee, and Ricardo Vilalta. 2019. Robust authorship verification with transfer learning. EasyChair Preprint no. 865.

Dainis Boumber, Yifan Zhang, and Arjun Mukherjee. 2018. Experiments with convolutional neural networks for multi-label authorship attribution. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Kevin Burton, Akshay Java, Ian Soboroff, et al. 2009. The icwsm 2009 spinn3r dataset. In *Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*.

Kevin Burton, Niels Kasch, and Ian Soboroff. 2011. The icwsm 2011 spinn3r dataset. In *Proceedings of the Annual Conference on Weblogs and Social Media (ICWSM 2011)*.

Deborah Cameron. 1996. Style policy and style politics: a neglected aspect of the language of the news. *Media, Culture & Society*, 18(2):315–333.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly. 2016. Stop clickbait: Detecting and preventing clickbaits in online news media. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 9–16.

Qian Chen, Ting He, and Rao Zhang. 2017. Deep learning based authorship identification.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

D. L. Davies and D. W. Bouldin. 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Paul Dickson and Robert Skole. 2012. *Journalese: A Dictionary for Deciphering the News*. Marion Street Press.

S. H. H. Ding, B. C. M. Fung, F. Iqbal, and W. K. Cheung. 2019. Learning stylometric representations for authorship analysis. *IEEE Transactions on Cybernetics*, 49(1):107–121.

Jade Goldstein-Stewart, Ransom Winder, and Roberta Sabin. 2009. Person identification from text and speech genre samples. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 336–344, Athens, Greece. Association for Computational Linguistics.

Shriya TP Gupta, Jajati Keshari Sahoo, and Rajendra Kumar Roul. 2019. Authorship identification using recurrent neural networks. In *Proceedings of the 2019 3rd International Conference on Information System and Data Mining*, ICISDM 2019, pages 133–137, New York, NY, USA. ACM.

Wynford Hicks. 2002. *Subediting for Journalists (Media Skills)*. Routledge.

David I. Holmes. 1998. The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, 13(3):111–117.

John Houvardas and Efstathios Stamatatos. 2006. N-gram feature selection for authorship identification. In *Artificial Intelligence: Methodology, Systems, and Applications*, pages 77–86, Berlin, Heidelberg. Springer Berlin Heidelberg.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Johannes Jasper, Philipp Berger, Patrick Hennig, and Christoph Meinel. 2018. Authorship verification on short text samples using stylometric embeddings. In *Analysis of Images, Social Networks and Texts*, pages 64–75, Cham. Springer International Publishing.

Jussi Karlgren. 2004. The wheres and whyfores for studying text genre computationally. In *Workshop on Style and Meaning in Languange, Art, Music and Design. National Conference on Artificial Intelligence*.

Mike Kestemont. 2014. Function words in authorship attribution. from black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 59–66, Gothenburg, Sweden. Association for Computational Linguistics.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page II–1188–II–1196. JMLR.org.

Rohith Menon and Yejin Choi. 2011. Domain independent authorship attribution without domain adaptation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 309–315, Hissar, Bulgaria. Association for Computational Linguistics.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Comput. Surv.*, 50(6):86:1–86:36.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540, New Orleans, Louisiana. Association for Computational Linguistics.

Jagadeesh Patchala and Raj Bhatnagar. 2018. Authorship attribution by consensus among multiple features. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2766–2777, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Tie-Yun Qian, Bing Liu, Qing Li, and Jianfeng Si. 2015. Review authorship attribution in a similarity space. *Journal of Computer Science and Technology*, 30(1):200–213.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, page 487–494, Arlington, Virginia, USA. AUAI Press.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. 2006. Effects of age and gender on blogging. In *Computational Approaches to Analyzing Weblogs - Papers from the AAAI Spring Symposium, Technical Report*, volume SS-06-03, pages 191–197.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. Authorship attribution with topic models. *Computational Linguistics*, 40(2):269–310.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Efstathios Stamatatos. 2017. 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*, pages 1138–1149, Valencia, Spain. Association for Computational Linguistics.

Efstathios Stamatatos. 2018. Masking topic-related information to enhance authorship attribution. *Journal of the Association for Information Science and Technology*, 69(3):461–473.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics*, pages 194–206, Cham. Springer International Publishing.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Andrew Weir. 2009. Article drop in english headlinese. *London: University College MA thesis*.

Min Yang, Xiaojun Chen, Wenting Tu, Ziyu Lu, Jia Zhu, and Qiang Qu. 2018. A topic drift model for authorship attribution. *Neurocomput.*, 273(C):133–140.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany. Association for Computational Linguistics.