# Writing Style Author Embedding Evaluation

**Enzo Terreau** **Antoine Gourru** **Julien Velcin**
Université de Lyon, Lyon 2, ERIC UR3083
{enzo.terreau, antoine.gourru, julien.velcin}@univ-lyon2.fr

## Introduction

- Learning authors representations from their textual production is widely used for multiple downstream tasks
- **Author embedding methods** are often built on top of either **Doc2Vec (Le and Mikolov, 2014)** or the **Transformer architecture (Devlin et al., 2019)**
- Most articles use either classification or authorship attribution as evaluation, which does not clearly measure the quality of the representation space, what it captures
- **We propose a novel evaluation framework of author embedding methods based on writing style**

## Author Embedding Evaluation

Extensions of Doc2Vec and BERT are the basic bricks of every author embedding method. Different strategies are used **to evaluate the quality of learnt representations:**

- **Link prediction**
- **Clustering** (evaluated with NMI)
- **Classification task**
- **Authorship attribution** (if document embedding can be inferred)

**These strategies are mostly centered on narrow downstream tasks.**
**None of them are strictly based on writing style.**

## Define a proxy of writing style

- Many works identify **the most relevant features to characterize writing style**
- Based on this, we choose a total of **301 stylistic features** (phonetic, syntactic, structural, …) **without any topic related information**
- We reach more than 80% of accuracy in authorship attribution on Project Gutenberg dataset (Table 2)
- These features are then aggregated among chosen axis, detailed Table 1

| Type of features | Examples | Number of features |
|---|---|---|
| Letters | Letter frequencies | 26 |
| Numbers | Numbers frequencies | 10 |
| Structural | Avg word length, Hapax Legomena, Syllable count, … | 9 |
| Punctuation | Punctuation sign frequencies | 16 |
| Function words | Function words (does, once, doing, …) frequencies | 174 |
| Tag | Pos tag frequencies | 43 |
| NER | Name Entity Recognition tag frequencies | 18 |
| Indexes | Complexity and readibility indexes | 7 |

Table 1: List of stylistic features selected and their categories. Frequencies are computed by sentence

| Authorship Attribution scores | | |
|---|---|---|
| **Numbers of authors** | **Accuracy** | **Coverage error** |
| 10 | 0.96 | 1.04 |
| 50 | 0.88 | 1.80 |
| 100 | 0.79 | 2.28 |

Table 2: Authorship attribution with logistic regression using only stylistic features.
With no topic information we reach 96% accuracy with 10 authors.
(Random sample of Project Gutenberg dataset)

## Proposed framework

**How to evaluate how well the embedding capture an author way of writing ?**

- Simple stylistic features are a good proxy of the authors' writing style
- They can easily be extracted from a corpus and aggregated by author

**Training a regression model using author embedding to predict these features allow to compare these representations in their ability to separate writing styles.**
Figure 1 shows the intuition behind this simple idea. (Code available here[2])

We use spacy tokenizers, POS-tagger and NER. NLTK stop words and CMU Dictionnary.
SVR with RBF Kernel (both quicker and better) for regression. (Code available here[2])

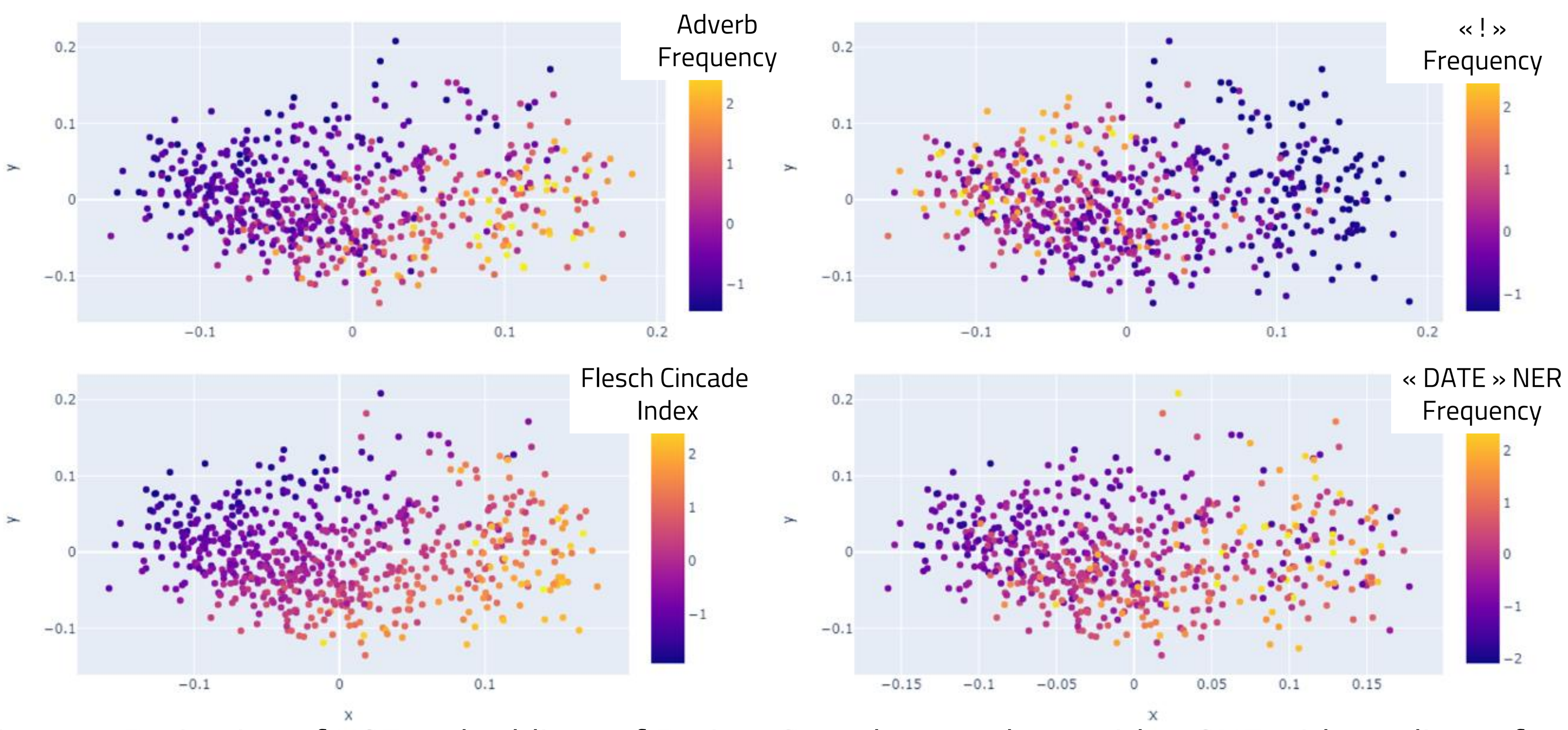


Figure 1: Projection of USE embeddings of Project Gutenberg authors with t-SNE with gradient of 4 selected stylistic features. Clear tendencies appear which motivates our method

[2]https://github.com/EnzoFleur/style_embedding_evaluation

## Datasets

- **Lyrics dataset[1]** (47 singers from various music genre, with at least 2,300 verses by author)
- Part of the **English Project Gutenberg** (664 authors with at least 10 books)
- Part of the **Blog Authorship Corpus** (500 bloggers with at least 50 blogposts)

[1]https://www.kaggle.com/paultimothymooney/poetry

## Competitors

To test our metric, we select two author embedding methods:

- **The Content Info Model** presented in Ganguly et al. (2016)
- **The annotated ngram based Doc2Vec** model of Maharjan et al. (2019)

We add two SOTA sentence embedding methods, which we extend to author embedding by averaging :

- **Sentence BERT** (Reimers and Gurevych, 2019)
- **Universal Sentence Encoder** (Cer et al. 2018), based on a Deep Averaging Network (DAN)

## Results

Results are presented below (Table 3, Figure 2):

- **Consistent author embeddings model obtain the worst MSE scores**
- **USE and SBERT can capture complex grammatical and linguistic notions**

BERT attention head naturally focuses on various linguistic phenomena in a sentence.

It is a surprise for the DAN version of USE. **It is the best model regarding our metric despite the BOW assumption made in the model:**

- **Syntactic treatment of sentences is not required to effectively represent them, even considering syntax and grammar**
- It is a huge improvement in terms of computation time

| Embedding | Letters | Numbers | Structural | Punctuation | Func. words | TAG | NER | Indexes |
|---|---|---|---|---|---|---|---|---|
| **USE** | **0.61 (0.27)** | **0.86 (0.09)** | **0.34 (0.18)** | 0.59 (0.26) | **0.65 (0.24)** | **0.45 (0.29)** | **0.65 (0.17)** | **0.27 (0.15)** |
| **Content Info** | 0.67 (0.22) | 0.87 (0.12) | 0.54 (0.18) | 0.67 (0.16) | 0.71 (0.19) | 0.65 (0.17) | 0.74 (0.13) | 0.50 (0.15) |
| **Ngram D2V** | 0.63 (0.20) | 0.88 (0.12) | 0.51 (0.20) | **0.58 (0.21)** | 0.68 (0.19) | 0.59 (0.19) | 0.71 (0.14) | 0.45 (0.15) |
| **SBERT** | 0.67 (0.27) | 0.90 (0.07) | 0.41 (0.19) | 0.62 (0.26) | 0.71 (0.21) | 0.51 (0.27) | 0.69 (0.18) | 0.32 (0.18) |

Average MSE Regression Score with standard deviation (SVR Model) on Gutenberg dataset

Table 3: MSE score (std in parenthesis) on the regression of stylistic features from author embedding on Gutenberg dataset using SVR. In bold, the best scores for each axis.

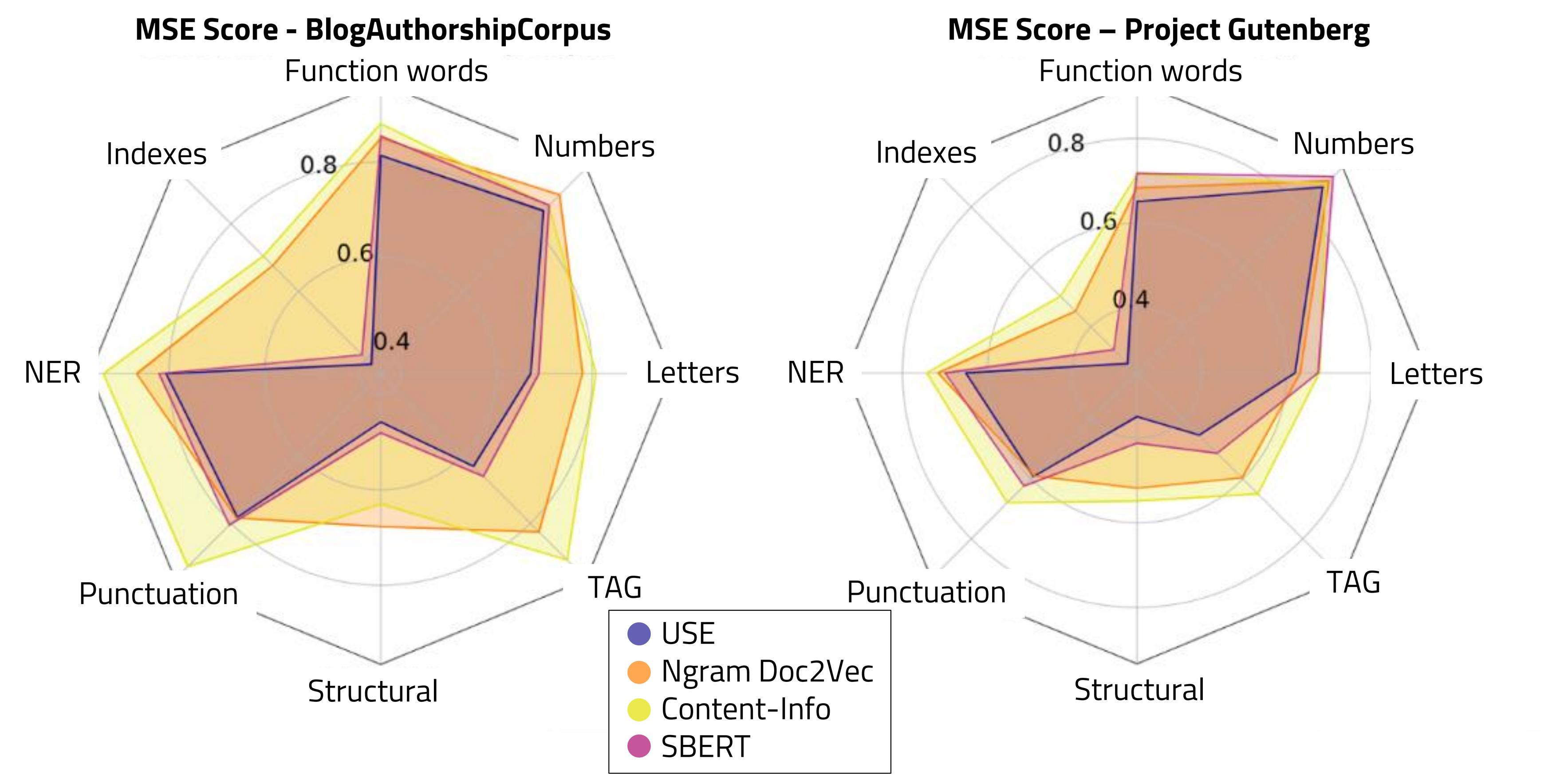


Figure 2: Spyder chart view of regression score on Gutenberg and Blog Authorship Corpus datasets

## Conclusion

- A **new evaluation framework for author embedding focusing on writing style**
- Based on the extraction of stylistic features, good proxy of an author way of writing
- Simple baselines outperform recent author embedding models in predicting most of those stylistic features
- **These SOTA sentence embedding models can capture complex linguistic notion thanks to multitask training on several big corpora**

## References

J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. *Bert: Pre-training of deep bidirectional transformers for language understanding.*
QV. Le and T. Mikolov. 2014. *Distributed representations of sentences and documents.*
S. Ganguly, M. Gupta, V. Varma, V. Pudi, et al. 2016. *Author2vec: Learning author representations by combining content and link information.*
D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y. Sung, B. Strope, and R. Kurzweil. 2018. *Universal sentence encoder.*
S. Maharjan, D. Mave, P. Shrestha, M. Montes-Y-Gómez, F. A. González, and T. Solorio. 2019. *Jointly learning author and annotated character N-gram embeddings: A case study in literary text.*
N. Reimers and I. Gurevych. 2019. *Sentence-bert:Sentence embeddings using siamese bert-networks*