SAURON: Leveraging Semantically Similar Utterances to Enhance Writing Style Embedding Models

---

Author:
Tim Koornstra
6435777

Supervisors:
Anna Wegmann
Dr. Dong Nguyen
Dr. Marijn Schraagen

A thesis submitted in fulfillment
of the requirements for the degree of
MSc. Artificial Intelligence
45 ECTS

Department of Information and Computing Sciences
Graduate School of Natural Sciences
College of Science

Universiteit Utrecht

Utrecht University
Utrecht, The Netherlands

June 2023

SAURON: Leveraging Semantically Similar Utterances to Enhance Writing Style Embedding Models, © June 2023

Author:
Tim Koornstra

Supervisors:
Anna Wegmann
Dr. Dong Nguyen
Dr. Marijn Schraagen

Institute:
Utrecht University, Utrecht, The Netherlands

CONTENTS

List of Figures . . . vi

List of Tables . . . ix

Abstract . . . xiii

1 INTRODUCTION . . . 1
1.1 Problem statement . . . 1
1.1.1 The components of written language . . . 1
1.1.2 Authorship Attribution and Verification . . . 2
1.1.3 Style modeling . . . 3
1.1.4 Challenges in style modeling . . . 4
1.2 Objectives . . . 5
1.3 Research questions . . . 5
1.4 Thesis outline . . . 6

2 RELATED WORK . . . 7
2.1 The definition of style . . . 7
2.2 Traditional approaches to modeling style . . . 9
2.2.1 Function words . . . 10
2.2.2 Burrow's delta . . . 11
2.2.3 Character n-grams . . . 12
2.3 Deep learning approaches . . . 13
2.3.1 Neural Networks . . . 13
2.3.2 Recurrent Neural Networks . . . 14
2.3.3 Transformer-based architectures . . . 15
2.4 Evaluation methods . . . 17

3 METHODS . . . 19
3.1 Dataset . . . 19
3.2 Training task . . . 22
3.2.1 Paraphrase mining . . . 25
3.2.2 Pairing utterances . . . 26
3.2.3 Fine-tuning pre-trained transformers . . . 27
3.3 Evaluation . . . 27
3.3.1 STEL framework . . . 27
3.3.2 Authorship Verification performance . . . 28
3.3.3 Baseline . . . 29

4 EVALUATION . . . 31
4.1 Semantic similarity analysis . . . 32
4.1.1 Score analysis . . . 32
4.1.2 Frequency distribution . . . 34
4.1.3 Paraphrase word count and score . . . 35
4.1.4 Comparison with Wegmann, Schraagen and Nguyen (2022) . . . 36
4.1.5 Conclusion . . . 39
4.2 Overview of experiments . . . 41
4.3 Initial experiments . . . 42
4.3.1 Hyperparameters . . . 43
4.3.2 Impact of random different-author sampling . . . 45
4.3.3 Summary of findings and baseline comparison . . . 47
4.4 Adjusting author sampling . . . 48
4.4.1 Results . . . 49
4.4.2 STEL-or-content analysis . . . 50
4.5 Experimentation with negative examples . . . 52
4.5.1 Results . . . 52
4.5.2 AV error analysis . . . 54
4.6 Randomly selecting utterances . . . 55
4.7 Non-uniform sampling of paraphrases . . . 57
4.8 Moving closer to Wegmann, Schraagen and Nguyen (2022) . . . 60
4.9 Main discussion . . . 62
4.9.1 (Contrastive) Authorship Verification . . . 62
4.9.2 STEL tasks . . . 63
4.9.3 Conclusion of proposed approaches . . . 65

5 CONCLUSION . . . 67
5.1 Research question . . . 67
5.1.1 Subquestion 1 . . . 67
5.1.2 Subquestion 2 . . . 68
5.1.3 Answer to main research question . . . 69
5.2 Limitations and future work . . . 70
5.2.1 Evaluation methods . . . 70
5.2.2 Interpretability . . . 71
5.2.3 The definition of style . . . 71
5.2.4 Semantic optimization . . . 72
5.2.5 Data diversity . . . 73
5.2.6 Time constraints . . . 73

A HYPERPARAMETER TUNING . . . 75
A.1 Loss function . . . 75
A.2 Learning rate . . . 75
A.3 Number of epochs . . . 76

A.4 Batch size 76
BIBLIOGRAPHY 76

LIST OF FIGURES

Figure 2.1 A STEL task instance. Anchor 1 (A1) and anchor 2 (A2) and the alternative sentences 1 (S1) and 2 (S2) are split along one of the proposed style dimensions: simple/complex. The task is to order S1 and S2 to match them to the same style as A1 and A2. In this figure, the utterances belonging to the same style have the same background color. Here, the correct order is thus S2-S1. 17

Figure 3.1 Two example text pairs for the AV task (these examples are made up). The task is to determine whether Utterances 1 (U1) and 2 (U2) were written by the same person or not. In the case of the first example, both utterances were written by the same author. This is thus a positive example. In the case of the second example, both utterances were written by a different author. This is thus a negative example. 23

Figure 3.2 An example text pair for the AV task where the texts are semantically similar (cosine similarity of 0.93) but stylistically different. In this case, both utterances were written by a different author. This is thus a negative example. 26

Figure 3.3 An example instance of a STEL-or-content task on the formal/informal dimension. In this case, the formal style of the Anchor sentence matches the style of Utterance 2, even though the content overlaps more with Utterance 1. 28

Figure 4.1 Histogram of paraphrase scores showing the distribution of scores across the dataset. The kernel density estimate (KDE) is also plotted to give a smooth estimate of the score distribution. The scores range from a minimum of 0.62 to a maximum of 1.00, with a median of 0.68. The 25th and 75th percentiles are 0.64 and 0.73, respectively. . . 32

Figure 4.2 Comparison of frequency distribution for the first 10,000 paraphrases in normal scale (left) and log scale (right), sorted by occurrences. The x-axis represents the index of the sorted paraphrases, while the y-axis represents the number of occurrences of each paraphrase. Both plots depict a rapid decrease in occurrence frequency, indicating a skewed distribution. The log scale on the right allows clearer visualization of the decline for less frequent paraphrases. The shaded area in both plots emphasizes the rate of decline. 35

Figure 4.3 Comparison of the top 15 most common paraphrases in the dataset before (left) and after (right) applying text cleaning processes such as converting to lower case, removing punctuation, and lemmatization. The total number of occurrences is represented on the x-axis in both plots. 36
Figure 4.4 Hexbin plot comparing the word count of the paraphrases and their corresponding cosine similarity scores. The x-axis represents the word count in the paraphrases, and the y-axis represents the score associated with the paraphrases. The color of the hexagons represents the count of paraphrase-score pairs that fall into the area, with darker colors indicating higher counts, plotted on a log scale. This plot visualizes the density and distribution of the data, as well as any potential correlations between paraphrase word count and the score assigned by the model. 37
Figure 4.5 Side-by-side comparison of the distribution of cosine similarity of the anchor-negative example pairs. The left plot shows the histogram for the sampling method that Wegmann, Schraagen and Nguyen (2022) used, while my sampling method is highlighted in the plot on the right. The kernel density estimate (KDE) is also plotted to give a smooth estimate of the score distribution. For a fair comparison, both plots have the same x and y limits. 38
Figure 4.6 Side-by-side comparison between the frequency distribution for the first 10,000 different-author examples used by Wegmann, Schraagen and Nguyen (2022) (left) and by me (right), sorted by occurrences. The x-axis represents the index of the sorted different-author examples, while the y-axis represents the number of occurrences of each different-author example. For a fair comparison, both plots have the same x and y limits. 39
Figure 4.7 Side-by-side comparison between hexbin plots comparing the word count of the different-author examples and their corresponding cosine similarity scores. This comparison is between the data used by Wegmann, Schraagen and Nguyen (2022) (left) and me (right). The x-axis represents the word count in the different-author examples, and the y-axis represents the score associated with the different-author examples. The color of the hexagons represents the count of example-score pairs that fall into the area, with darker colors indicating higher counts, plotted on a log scale. This plot visualizes the density and distribution of the data, as well as any potential correlations between example word count and the score assigned by the model. For a fair comparison, both plots have the same x and y limits. 40
Figure 4.8 A flowchart summarizing the various different experiments and resulting models presented in the following sections of this thesis. . . . 41

Figure 4.9 One of the STEL-or-content task instances on the formal/informal dimension. In this case, the ground truth is that the Anchor and Utterance 2 are written in the same style. My model assigns a cosine similarity score of 0.363 to the Anchor and U2, and a score of 0.707 to the Anchor and U1. 51
Figure 4.10 Comparison of input examples for the Contrastive Authorship Verification task with the old (left) and new (right) sampling approach. In both scenarios, Utterance 1 and Utterance 2 correspond to the same-author and different-author examples, respectively. 52
Figure 4.11 Comparison of confusion matrices of the results on the Authorship Verification task with the 100% (left) and 0% (right) semantically similar different-author examples test sets. 55
Figure 4.12 The cosine similarity score distribution before (left) and after (right) applying the new weighted sampling method. The kernel density estimate (KDE) is also plotted to give a smooth estimate of the score distribution. For a fair comparison, both plots have the same x and y limits. 58

LIST OF TABLES

Table 3.1 Comparison of utterances and authors used by relevant related works, broken down by data source. In the case of this thesis, the listed figures represent data quantities before the implementation of any pre-processing or selection steps. The final quantity of data utilized, post-processing, is notably less and will be detailed in the ensuing section. 20

Table 3.2 Summary of preprocessing steps and their impact on the dataset, showing the number of removed utterances and authors, percentage of data removed, and remaining data after each step, along with the ratio of remaining utterances per author. Steps that do not remove any utterances or authors were left out for brevity. 22

Table 3.3 Overview of the number of utterances and number of authors for the dataset before and after pre-processing. 22

Table 4.1 Examples of the highest scoring paraphrases from the data. These examples represent paraphrases that are either identical or very similar to the anchor sentences, leading to high cosine similarity scores. 33

Table 4.2 This table presents 5 of the lowest scoring anchor-paraphrase pairs that were used. These examples depict substantial variations in terms of length, wording, context, or semantic meaning compared to the anchor sentences, which result in low cosine similarity scores. 34

Table 4.3 Table comparing the performance of the RoBERTa base model, the model from Wegmann, Schraagen and Nguyen (2022), and all models presented in this thesis on the AV and CAV tasks using 0% and 100% semantically similar negative examples, as well as the STEL framework. The Formal, Complex, Number substitution and Contraction task columns are subdivided into Original (standard STEL dataset) and o-c (STEL-or-content dataset) subcolumns. The values with the highest accuracy in each column are reported in **bold**. The underlined values correspond to the highest accuracy in each column for the models that I present throughout this chapter. 43

Table 4.4 The table above presents the experiments' results on the impact of random different-author sampling on model performance, with accuracies displayed as percentages. Columns for the AV Task and CAV Task show results for 0% and 100% semantically similar conditions, respectively. The Formal, Complex, Number substitution and Contraction task columns are subdivided into Original (standard STEL dataset) and o-c (STEL-or-content dataset) subcolumns. The values with the highest accuracy in each column are reported in bold. 46

Table 4.5 Table comparing the performance of the RoBERTa base model, the model from Wegmann, Schraagen and Nguyen (2022), and the proposed model ("Initial model") on the AV and CAV tasks using 0% and 100% semantically similar different-author examples, as well as the STEL framework. The table shows the results for both the original tasks and the STEL-or-content (o-c) tasks. The values with the highest accuracy in each column are reported in bold. 47

Table 4.6 Table comparing the performance of the model from Wegmann, Schraagen and Nguyen (2022) and my initial model on the AV and CAV tasks, using the test sets from their paper. This testing task features their conversation-level content control. The values with the highest accuracy in each column are reported in bold. 48

Table 4.7 Table comparing the performance of the RoBERTa base model, the model from Wegmann, Schraagen and Nguyen (2022), the model from § 4.3.3 ("Initial"), and the new proposed model ("1 utterance authors") on the AV and CAV tasks using 0% and 100% semantically similar different-author examples, as well as the STEL framework. The table shows the results for both the original tasks and the STEL-or-content (o-c) tasks. The values with the highest accuracy in each column are reported in bold. 49

Table 4.8 Table comparing the performance of the RoBERTa base model, the model from Wegmann, Schraagen and Nguyen (2022), the previously proposed model ("1 utterance authors"), and the new proposed model ("Positive-negative") on the AV and CAV tasks using 0% and 100% semantically similar negative examples, as well as the STEL framework. The table shows the results for both the original tasks and the STEL-or-content (o-c) tasks. The values with the highest accuracy in each column are reported in bold. 53

Table 4.9 Table comparing the performance of the RoBERTa base model, the model from Wegmann, Schraagen and Nguyen (2022), the previously proposed model ("1 utterance authors"), and the new proposed model ("Random sampling") on the AV and CAV tasks using 0% and 100% semantically similar different-author examples, as well as the STEL framework. The table shows the results for both the original tasks and the STEL-or-content (o-c) tasks. The values with the highest accuracy in each column are reported in bold. . . . 56
Table 4.10 Table comparing the performance of the RoBERTa base model, the model from Wegmann, Schraagen and Nguyen (2022), the previously proposed model ("Random sampling"), and the new proposed model ("Non-uniform") on the AV and CAV tasks using 0% and 100% semantically similar different-author examples, as well as the STEL framework. The table shows the results for both the original tasks and the STEL-or-content (o-c) tasks. The values with the highest accuracy in each column are reported in bold. 58
Table 4.11 Table comparing the performance of the RoBERTa base model, the three contrastive models from Wegmann, Schraagen and Nguyen (2022), and the proposed model ("Non-uniform") on the STEL framework. The table shows the results for both the original tasks and the STEL-or-content (o-c) tasks. The values with the highest accuracy in each column are reported in bold. 59
Table 4.12 Table comparing the performance of the RoBERTa base model, the model from Wegmann, Schraagen and Nguyen (2022), the previously proposed model ("Non-uniform"), and the new proposed model ("SAURON") on the AV and CAV tasks using 0% and 100% semantically similar different-author examples, as well as the STEL framework. The table shows the results for both the original tasks and the STEL-or-content (o-c) tasks. The values with the highest accuracy in each column are reported in bold. 61
Table 4.13 Table comparing the performance of the RoBERTa base model, the model from Wegmann, Schraagen and Nguyen (2022), and all models presented in this thesis on the AV and CAV tasks using 0% and 100% semantically similar negative examples, as well as the STEL framework. The Formal, Complex, Number substitution and Contraction task columns are subdivided into Original (standard STEL dataset) and o-c (STEL-or-content dataset) subcolumns. The values with the highest accuracy in each column are reported in **bold**. The underlined values correspond to the highest accuracy in each column for the models that I present throughout this chapter. **Note:** This table is the same as Table 4.3. 63

Table A.1 The table above presents the results of the impact of two different loss functions on the test tasks. Columns for the AV Task and CAV Task show results for 0% and 100% semantically similar conditions, respectively. The Formal, Complex, Number substitution, and Contraction task columns are subdivided into Original (standard STEL task) and o-c (STEL-or-content task) subcolumns. The values with the highest accuracy in each column are reported in bold. 75
Table A.2 The table above presents the results of the impact of various learning rates on the test tasks. Columns for the AV Task and CAV Task show results for 0% and 100% semantically similar conditions, respectively. The Formal, Complex, Number substitution, and Contraction task columns are subdivided into Original (standard STEL task) and o-c (STEL-or-content task) subcolumns. The values with the highest accuracy in each column are reported in bold. 75
Table A.3 The table above presents the results of the impact of the number of training epochs on the test tasks. Columns for the AV Task and CAV Task show results for 0% and 100% semantically similar conditions, respectively. The Formal, Complex, Number substitution, and Contraction task columns are subdivided into Original (standard STEL task) and o-c (STEL-or-content task) subcolumns. The values with the highest accuracy in each column are reported in bold. 76
Table A.4 The table above presents the results of the impact of various batch sizes. Columns for the AV Task and CAV Task show results for 0% and 100% semantically similar conditions, respectively. The Formal, Complex, Number substitution, and Contraction task columns are subdivided into Original (standard STEL task) and o-c (STEL-or-content task) subcolumns. The values with the highest accuracy in each column are reported in bold. 76

ABSTRACT

This thesis investigates the potential for enhancing transformer-based models, widely used in Natural Language Processing (NLP), for the task of writing style representation. I propose a novel approach wherein a RoBERTa model (Liu et al., 2019) is trained on the Contrastive Authorship Verification (CAV) task using semantically similar utterances. These are pairs of utterances that encapsulate the same semantic information but differ in their stylistic expression. This methodology encourages the model to concentrate more on style rather than content, fostering a more discerning representation of stylistic nuances. The training data comprised a broad array of conversations from the online platform Reddit, providing a wide representation of authorship and topics.

To assess the performance of the models, the STyle EvaLuation (STEL) framework (Wegmann and Nguyen, 2021) was utilized. The results of the STEL evaluation helped ascertain the models' ability to accurately capture writing style and delineate the impact of introducing semantically similar pairings.

While incorporating semantically similar utterances greatly improved performance over models without any form of content control, it was discovered that relying solely on semantically similar utterances was not the most efficient approach. Instead, the findings suggested that a combination of this technique with conversation-based sampling of examples could further enhance the models' performance. Additionally, the research underlined various effective strategies for preparing input data, such as maintaining diversity in authorship and topics.

The final model, coined as the SAURON (Stylistic AUthorship RepresentatiON) model, considerably improved upon previous iterations. This advancement contributes to the advancement of style-content disentanglement tasks and paves the way for more nuanced and robust style representations.

The code developed for this project is freely available on GitHub[1] and the trained SAURON model can be accessed on the Huggingface Hub[2]. These resources are provided for public use, further development, and to encourage reproducibility and transparency in research.

1 https://github.com/TimKoornstra/SAURON
2 https://huggingface.co/TimKoornstra/SAURON

1 INTRODUCTION

1.1 PROBLEM STATEMENT

Language plays a crucial role in our daily lives, serving as a means of communication and self-expression. Specifically, the written dimension of natural language is becoming increasingly important, as evidenced by the rising global literacy rates and internet connectivity. In 2016, 86% of the global population above the age of 15 were able to read and write (vs. 42% in 1961) (Roser and Ortiz-Ospina, 2016), and by 2022, 63% of the world's population was connected to the internet (vs. 41.5% in 2015) (Clement, 2022). As a result, the written dimension of natural language has become an important tool for sharing ideas, emotions, and information, as well as for building and maintaining social connections. With the rise of social media and other online platforms, the volume of written language has grown exponentially, making it a crucial area of study for understanding human communication and expression.

1.1.1 *The components of written language*

Written language is a complex medium of communication that can be divided into different parts. Two of the main components traditionally studied in linguistic analysis are syntax, which refers to how words are arranged in a sentence, and semantics, which refers to the meaning conveyed by those words. Style, on the other hand, is an aspect that is independent of meaning and can be analyzed separately, encompassing elements such as word choice, sentence length, and tone. It's worth noting that this distinction is not exhaustive and other elements such as cultural and social context, the author's personality, and the intended audience can also play a role (Funkhouser and Maccoby, 1973; Wolfradt and Pretz, 2001).

One way to understand style in written language is as the set of linguistic choices an author makes to convey a message in a distinct way or express a certain tone (Hacker, 1994; Ross-Larson, 1999). According to Sebranek, Kemper and Meyer (2006), these choices can include but are not limited to, basic elements such as grammar and punctuation, as well as more intricate choices like sentence structure, vocabulary, and paragraph organization. These choices can greatly impact how a message is received by the reader, even if the underlying message is the same. For example, the style of a formal business report will typically be more structured, precise, and professional than that of a casual

email to a friend, which may use informal language and a more conversational tone. While the content of the two messages could theoretically be the same, their different styles convey different attitudes and expectations about the message and the intended audience. The manner in which language is used in a text can thus greatly influence audience perceptions of its quality and persuasiveness. A study conducted by Chartprasert (1993) found that subjects rated authors with a wordy and difficult-to-understand writing style to have higher expertise than authors who wrote in a simple style. Other research has found that language style has a significant impact on the perceived usefulness of online reviews (Liu, Xie and Zhang, 2019; Yang, Zhou and Chen, 2021), as well as increased conversion rates on websites (Ludwig et al., 2013). Van den Besselaar and Mom (2022) also found that the use of complex language (e.g., longer text and longer sentences), alongside technical content (e.g., less common words) on research grant applications increases the chances of acquiring them.

Additionally, the way in which we express ourselves in written language is also important since the nuances and subtleties of spoken language, such as intonation and stress, are not as easily conveyed, which poses different challenges and difficulties in effectively conveying the intended meaning. Although other factors than style play a part here, a great example of such challenges is sarcasm (Filik, Hunter and Leuthold, 2015), which can be difficult to recognize in written language without additional context. It often requires lexical or pragmatic stylistic features, such as the ending a comment or post with "/s" on Reddit (Emerson, 2022) or the use of emoticons (Thompson and Filik, 2016), to contextualize or motivate an utterance (Skovholt, Grønning and Kankaanranta, 2014). These features are a part of written language, and the way they are used can greatly impact how the intended meaning is conveyed.

1.1.2 *Authorship Attribution and Verification*

As previously established, the writing style of an author holds great significance in written language. In fact, it can be so distinct that it can serve as a unique fingerprint that can be used to help identify unknown authors (Bergs, 2015). This idea is utilized in the fields of **Authorship Attribution** (AA) and **Authorship Verification** (AV). The goal of the former task is to determine the identity of the author of a given piece of text. This is sometimes done as a means of identifying the source of a document or determining whether a particular individual wrote a given piece of writing (Stamatatos, 2009). The goal of Authorship Verification, on the other hand, is to certify the author of a text. The task takes as input a pair of texts and outputs a decision of whether both texts were written by the same author. This is often used as a means of verifying the authenticity of a document or establishing the identity of an individual (Stamatatos, 2016). So, Authorship Attribution aims to identify the author of a given text, while Authorship Verification aims to confirm whether a specific individual is the author of a given text. Models trained on these tasks should not only be able to determine the identity of the author and certify the authorship of a text,

but should also be able to distinguish different individuals based on factors such as the content of the text, and their writing style. Both tasks also provide (i) a means of testing and evaluating the effectiveness of the models developed, by assessing their performance on unseen data, and (ii) provide a way to check the robustness and generalizability of the models, which are important factors in evaluating the performance of the models (Goodfellow, Shlens and Szegedy, 2015). Some examples of this are: determining the author of misinformation (Buda and Bolonyai, 2020), verifying the authenticity of tweets (Theophilo, Giot and Rocha, 2021), establishing the credibility of a source (Choi and Lim, 2019), or detecting whether a text was truly generated by a human. Long-established research fields such as humanities and history could use these techniques to determine who authored a document (Ouamour and Sayoud, 2013), and whether they were influenced or helped by someone when writing it (Zhao and Zobel, 2007). The police could employ AA or AV to determine who wrote anonymous messages when tapping criminal phones, or when analyzing extremist forums, for instance (Chaski, 2005).

1.1.3 *Style modeling*

One way to verify the authorship of a document is through the analysis of writing style. This aspect of an individual's writing can be captured computationally through **style representation** or **style modeling**. Style representation involves extracting and representing the characteristics of an author's writing style in a numerical or symbolic form, using features such as vocabulary, grammar, sentence structure, and other patterns (e.g., Holmes and Forsyth (1995) and Kestemont et al. (2012)). These features are then used to create a model that can represent and distinguish the writing style of one author from another.
Traditionally, manual feature engineering has been the most prominent technique for researchers to describe style, but this method comes with the drawback of being time-consuming and data-intensive (Amir et al., 2016). Another challenge with this approach is that style is sometimes very subtle and hard to manually craft rules for. Even small variations in the frequency distribution of certain words or linguistic patterns, for example, may not be captured by traditional manual feature engineering methods, which can lead to an incomplete representation of an author's writing style (Rudman, 1997). Furthermore, manual feature engineering can sometimes involve improving accuracy iteratively by finding new features that distinguish a specific author from other authors (Koppel and Schler, 2004). Although this technique succeeds in verifying the authorship of specific authors, it does not do well when the goal is to verify authors that are not within the training data, since there are no rules available for those authors yet. In other words: although the style of individual writers can be expressed as a set of crafted rules through this method, the creation of a general representation of writing style might not be possible this way.
Thanks to advances in computing power and machine learning in the 2000s and 2010s, it has become easier to computationally process text and apply deep neural network-style methods in natural language processing (NLP). Despite that overall there has been great progress in NLP, there has been less focus on the area of style representation because state-

of-the-art linguistic representation methods (e.g., Sentence-BERT (Reimers and Gurevych, 2019)) tend to focus on the semantic rather than the stylistic embeddings since these methods are primarily designed for tasks such as text classification, machine translation, and question answering, which require a representation of the meaning of text rather than its style (Devlin et al., 2019; Vig and Belinkov, 2019).

1.1.4 *Challenges in style modeling*

One of the main challenges in automated style modeling is the separation of style from content. Because style and content are intertwined (Kaplan, 1968), it is difficult to separate the two. The task of separating style from content is complicated by the fact that many modern approaches do not explicitly differentiate between the two. These models often rely on complex, high-dimensional representations of text, which can capture both content and style information in a single representation. This makes it difficult to disentangle the two and extract only the style information. Additionally, style is often intertwined with the meaning of the text, and so separating the two can result in a loss of information.
This problem can be mitigated by **controlling for content**, which refers to ensuring that the stylistic variations in the text being analyzed are not due to variations in the underlying content, but rather the choice of language used by the author. An example of controlling for content in automated style modeling is the use of text distortion to mask words that do not occur often in a language (Stamatatos, 2017; Stamatatos, 2018). The idea here is that these words tend to be specific to a topic, making them less likely to carry information about style and more likely to carry information about the content of the text. Although there are some good results with approaches to the AV and AA tasks that do not control for content (as listed in Stamatatos (2016), for example), verifying the authorship of a document becomes difficult when an author writes about something different than they would usually write about (Bischoff et al., 2020). This is because an author's choice of words and phrases may be influenced by the topic they are writing about, which could lead to inaccuracies in the Authorship Verification and Attribution tasks if the model is not able to separate the author's writing style from the content-specific semantics.

The recent innovation of transformer-based architectures such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019) has led to numerous breakthroughs in NLP and language modeling. Although there has been some research that has tried to model stylistic vector representations using transformer-based architectures, such as the work by Rivera-Soto et al. (2021) and Wegmann, Schraagen and Nguyen (2022), the models developed in these studies have shown limitations in terms of cross-domain transferability and the ability to effectively disentangle content and style in the representations.

1.2 OBJECTIVES

In this project, the primary goal is to enhance the representation of writing styles by leveraging transformer-based approaches, building upon previous work in the field. The principal strategy for overcoming the limitations of previous work is to train models on the AV task with similar utterances that convey the same meaning, which might mitigate bias towards the content of the utterances and improve generalization across different domains. This is because, when training on semantically similar utterances, the model is exposed to variations in writing style while being presented with similar content. The idea is thus that since the content of the utterance is almost the same, the way in which it is said must be different. This allows the model to focus on learning the nuances of writing style to verify the authorship of a document, rather than being distracted by variations in content. This is an interesting area of research that has yet to be fully explored and has the potential to improve the performance of stylistic vector representations.
The effectiveness of the proposed approach will be evaluated using the STEL framework as introduced by Wegmann and Nguyen (2021), which allows for the ability to determine to which extent a model can disentangle style from content. The results from this novel approach will be benchmarked against a key baseline: Wegmann, Schraagen and Nguyen (2022). This baseline was chosen based on its utilization of transformer-based architectures for style modeling and the fact that their model is openly available. Furthermore, this work was selected as a benchmark due to its demonstrated effectiveness at disentangling style from content and its use of a content-control approach similar to the one proposed in this thesis.

A more comprehensive description of the project setup and research methods can be found in Chapter 3. While certain potential solutions such as leveraging data from diverse social media sources to increase model robustness and generalization performance have been discussed in previous works, this thesis will primarily focus on the strategy of training models on semantically similar but stylistically diverse utterances. Other solutions, though promising, are outside the scope of this work.

1.3 RESEARCH QUESTIONS

This thesis aims to answer the following research questions and sub-questions as comprehensibly as possible:

RQ1. *How does incorporating semantically similar utterances affect the performance of transformer-based approaches for writing style representation?*

- SQ1. How does the use of semantically similar utterances compare to using other types of control data (e.g. unrelated sentences, sentences from the same conversation)?
- SQ2. What are the most effective sampling techniques for preparing the input data?

1.4 THESIS OUTLINE

In this thesis, I will first conduct a literary study on the topic and provide the motivation for the methods I have chosen for my research in Chapter 2. Then, in Chapter 3, I will describe the methods I will use in my research. This will include the process of collecting and pre-processing data, developing and training models, and evaluating the proposed approach. In Chapter 4, I will present an iterative evaluation of the proposed approach, including the use of baseline models, the STEL framework, and performance metrics. Finally, in Chapter 5, I will provide a summary of the main findings of the thesis, discuss their implications, and suggest potential future directions for research in this area.

2

RELATED WORK

In this chapter, I will be analyzing some related works that are relevant to the topic of this thesis and that will assist me in answering the research questions and sub-questions. In Section § 2.1, I will first look at the linguistic debate about style to find a definition that I will use throughout this thesis. Then, in Section § 2.2 I will look into the traditional approaches to modeling style, their shortcomings, and what we can learn from them. Section § 2.3 will describe modern deep learning approaches and I will weigh their advantages and disadvantages. The last section in this chapter - Section § 2.4 - describes an evaluation method that can be used for the Authorship Verification task.

2.1 THE DEFINITION OF STYLE

What defines style? This is not a trivial question to answer, and there has indeed been ample debate on this topic. However, in order to answer my research questions and conduct my research, I will characterize *writing* style and present a rough definition that I shall use throughout this thesis.

Style, which is not limited to language, can be found across various fields and is often distinguished between individual and group styles. Individual style can refer to the unique characteristics and features that distinguish an individual's style from others, while group style can refer to the shared characteristics and features that are common among members of a group (Kent, 1986). For example, in architecture, the Gothic style is characterized by the use of pointed arches, rib vaults, and flying buttresses, and is common among buildings built in the Middle Ages (Fraser, 2018). In music, the blues genre is characterized by the use of a 12-bar chord progression and the use of the blues scale and is common among blues musicians (Wikipedia, 2023).
Chan (1994) found that not only can architectural group style be discovered by recognizing common features present in buildings, but individual style is also represented. In music, individual style is often characterized by the use of particular instruments, melodies, and harmonies (Juslin and Sloboda, 2001). Jazz musicians such as Miles Davis and John Coltrane are known for their unique approaches to improvisation and the use of specific melodies and harmonies, while classical composers such as Ludwig van Beethoven and Wolfgang Amadeus Mozart are known for their distinctive styles characterized by specific instrumentation and formal structure.

We can see that the recognition of both individual and group styles is usually achieved by recognizing common features. Holmes (1994) thus describes style as a "set of measurable patterns which may be unique to an author", and the Cambridge Dictionary defines it as "a way of doing something, especially one that is typical of a person, group of people, place, or period" (Cambridge Dictionary (2023)). Interestingly, Biber and Conrad (2009) separate the way of expressing oneself into three different concepts: register, genre, and style. They define genre as a category of texts that share similar communicative purposes and social contexts, such as news articles, scientific research articles, and fiction novels. They define register as a variation of language associated with a particular subject matter or situational context, such as the language used in a legal document versus casual conversation. Lastly, they define style as the choices a writer makes in terms of vocabulary, grammar, and other linguistic features that are associated with their individual writing habits or the conventions of a particular genre or register.
From all these definitions it becomes apparent that style is an important factor in distinguishing one's work from that of others. This is especially true in the realm of writing, where individual style allows writers to convey their message in a unique and personal way through the use of specific words, literary devices, structure, and tone (Sebranek, Kemper and Meyer, 2006) — an idea that is exploited in forensic linguistics by means of "linguistic fingerprinting" (Coulthard, 2014). In contrast, some scholars in the field of sociolinguistics understand group style as a category of written or spoken communication that is characterized by shared features and conventions (Wardhaugh and Fuller, 2021). These conventions may be associated with particular fields or disciplines and serve as a means of creating and communicating within a specific group or style. The combination of both individual and group style is important for the creation of unique writing. Great literary examples of this are the writers F. Scott Fitzgerald (Keshmiri and Mahdikhani, 2015) and Ernest Hemingway (Xie, 2008); both managed to create a distinctly individual style within a well-established group style to create unique works. In fact, the ability to effectively weigh the conventions of a particular group style against the personal touch of individual style is essential for the tasks of Authorship Attribution and Authorship Verification. Thus, by considering both the group style and the individual style of a piece of writing, it is possible to accurately attribute the work to a particular author or verify that an author is the true creator of a piece of writing.

In this thesis, **style** refers to the distinct patterns and features of language use, such as word choice, sentence structure, and use of punctuation, that can be used to identify and distinguish the writing of a particular author from others. It is important to know that these patterns and features may be influenced by the conventions and trends within a specific group or community, but do not include the content of the writing.

2.2 TRADITIONAL APPROACHES TO MODELING STYLE

Stylometry, which involves the examination and evaluation of the language and literary techniques used in written texts, has a long history. Early efforts in this field often relied on manual analysis of specific features, such as word choice and punctuation, without considering other criteria such as parts of speech and sentence length, which are considered to be more objective and less subject to interpretation. Some notable examples of this type of approach include the evaluations of writers by Addison and Steele (1711) in their 18th-century periodical, The Spectator. In this publication, they used their own personal criteria and observations to assess the style of various writers. This likely included elements such as word choice, sentence structure, and literary techniques, but it is not clear exactly what criteria they used as it is not specified in the historical record. In the 20th century, writers such as William Strunk Jr. and E.B. White took a more systematic approach, writing "The Elements of Style," a guide to effective writing that covers topics such as word choice, sentence structure, and the use of figurative language (Strunk and White, 1972). Scholars such as Cleanth Brooks and Robert Penn Warren also used close readings of literature to evaluate the style of writers in the 20th century, focusing on the use of rhetorical devices and the structure of poetry (Brooks, 1947; Warren, 1952).

The issue with these methods is that they are widely divergent and they are not reproducible, as they rely on personal criteria and observations which can vary greatly from person to person. This makes it difficult to replicate results or compare evaluations of different writers. Furthermore, the wide divergence in these approaches makes it challenging to compare and replicate the results of different studies, as the criteria and methods used can vary greatly. This lack of standardization can lead to inconsistent and unreliable results, making it difficult to draw meaningful conclusions about the writing styles being examined. Additionally, the wide divergence in these approaches also makes it hard to establish a consensus or understanding about what constitutes good writing or effective stylistic analysis. The issue of reproducibility and standardization can pose a challenge in Authorship Attribution, as it can be difficult to determine the true author of a piece of writing (Stamatatos, 2009), as is seen with historical examples such as The Federalist Papers (Adair, 1944) — a series of 85 essays written in the late 1700s to promote the ratification of the United States Constitution. Without a clear and objective evaluation method, it becomes difficult to assess the accuracy and reliability of the results.

The work of Mosteller and Wallace (1963), however, marked a turning point, as they introduced a more systematic and reproducible quantitative approach based on function words, which led to the development of another widely used method: character n-grams. In this section, I will review the early history of style analysis and the contributions of Mosteller and Wallace, and discuss how these approaches have been used in subsequent research. I will also delve deeper into the use of function words and character n-grams for style analysis, and discuss their strengths and limitations.

2.2.1 *Function words*

Function words are words that indicate the grammatical relationship between other words or phrases in a sentence. They include words such as articles (e.g., "a," "an," "the"), pronouns (e.g., "I," "you," "he"), conjunctions (e.g., "and," "but," "or"), and prepositions (e.g., "in," "on," "under"). In their pioneering research on Authorship Attribution, Mosteller and Wallace (1963) found that function words are more stable and less likely to vary across different writing contexts compared to content words (i.e., words that convey meaningful content). This stability in function words across different writing contexts makes them a valuable tool for Authorship Attribution, as they can provide a more consistent measurement for comparison. Mosteller and Wallace developed a statistical model that used the distribution of function words in texts to predict the likelihood that a given text was written by a particular author. They applied this model to a case of disputed authorship, specifically the authorship of some of the Federalist Papers, and found that it was able to correctly[1] identify the true author in many instances, despite the lack of verified labels in the dataset.

In the early days of computational stylistics, Damerau (1975) confirmed Mosteller and Wallace's work and proposed that function words, along with other linguistic features, could indeed be used to identify the authorship of a text, albeit that he deemed "the satisfactoriness of function words [...] to be doubtful". He recommended that the search for "minor encoding habits" (i.e., the small and often unconscious choices that an author makes when writing a text), such as the use of function words, as indicators of style and authorship should be pursued more vigorously in other areas. On top of that, as pointed out by Damerau and Mandelbrot (1973), the clustering of high-frequency functional words was computationally infeasible at the time. Due to these difficulties, it took many years for function words to actually become popular.
Because of a massive improvement in computational power in the last two decades, several recent studies have used function words to analyze authorship and style. For instance, Argamon et al. (2003) found that different authors have distinct function word signatures and used this in combination with other linguistic features to identify the authorship of texts, such as parts of speech, to identify the authorship of texts. Pennebaker, Mehl and Niederhoffer (2003) used function words in personal emails as stylistic features to predict the writer's extraversion and emotional state. Pennebaker (2011a) has also proposed that function words can be used to study the psychological underpinnings of writing style in his book "The Secret Life of Pronouns" and has shown how function words can be used to identify aggressive intent and language style (Pennebaker, 2011b). Thus, function words have been used for both identifying authorship and studying the psychological underpinnings of writing style and these studies demonstrate the potential of function words in understanding and analyzing writing style.

1 *Correct* in this case is defined as a prediction that aligns with the consensus of experts on the authorship of a given text.

The use of function words to describe writing style can be problematic for a number of reasons, however. Firstly, a number of studies have shown that function words, and especially personal pronouns, do not solely reflect an author's writing style, but also correlate with other factors such as narrative perspective, an author's gender, or even a text's topic. For example, Paisley (1964) found that not all acknowledged function words are free of context and that certain words such as "I" and "we" were more topic-oriented rather than style-oriented. Secondly, it can be difficult to determine which function words are the "best" to use for style analysis. Damerau and Mandelbrot (1973) found that different function words can have different degrees of usefulness for different types of writing, making it challenging to establish a consistent method for style analysis. A third reason is that the highly reductionistic nature of function words also seems unsatisfying as they rarely give a good insight into underlying stylistic issues (Argamon and Levitan, 2005). Another point of concern raised is that the restriction to function words for stylometric research seems sub-optimal for languages that make less use of function words (Rybicki and Eder, 2011). Finally, the use of function words as a sole indicator of writing style can also be problematic as it limits the scope of stylistic analysis. There are certain style features that cannot be captured by examining function words alone, such as spelling variations, use of emojis, and other non-verbal forms of expression that are prevalent in social media communication. These features, which are not captured by function words, can be crucial for understanding the unique writing style of an author, especially in the context of social media. Therefore, relying solely on function words to describe writing style may not provide a comprehensive understanding of the nuances of an author's style.

2.2.2 *Burrow's delta*

In addition to the use of function words, another common, related method in stylometry is the application of Burrows' Delta (Burrows, 2002). Burrows' Delta is a measure of the difference between the frequency of words in a given text and the frequency of those same words in a reference corpus. By comparing an author's use of words to the general patterns of language use in a reference corpus, it is possible to identify the distinctive vocabulary and style of an author. Burrows' Delta has been used successfully in a number of studies, including identifying the authorship of disputed texts and tracing the evolution of an author's style over time (Hoover, 2004; Hoover, 2012). However, Burrows' Delta has some limitations as well. For example, it may not be effective in identifying certain types of stylistic features that do not involve word frequency, such as sentence structure or use of metaphors. Furthermore, it may be less effective in identifying the style of authors who use a limited vocabulary, or in languages with a smaller corpus of reference texts (Juola, 2008) Despite these limitations, Burrows' Delta remains a useful tool in stylometry and can be used in conjunction with other methods, such as the analysis of function words, to gain a more complete understanding of an author's style.

2.2.3 *Character n-grams*

Another frequency-based approach that was popularized by the success of Mosteller and Wallace (1963) was the use of character n-grams. Character n-grams are contiguous sequences of n characters in a text. For example, the character n-grams for the word "hello" include "hel," "ell," and "llo" (for $n = 3$). The concept of n-grams has a long history, with roots dating back to a few years after the end of World War II. One of the earliest uses of n-grams was in the development of language models, where they were used to predict the likelihood of a sequence of words in a text. Claude Shannon introduced the use of n-grams for this purpose in his 1951 paper, "Prediction and Entropy of Printed English" (Shannon, 1951). As a variation on regular n-grams, character n-grams can be used to analyze the style of a text by looking at the frequency of different n-grams in the text. By analyzing the frequency of different character n-grams in a text, it is possible to gain insights into the writing style of the author and identify patterns that may be unique to a particular individual. In the study by Kjell, Woods and Frieder (1994), character n-grams were used to determine the authorship of 12 unattributed papers in the Federalist Papers. The authors utilized visualization techniques to help organize the vast amount of data generated in computational studies of literary style. These techniques were demonstrated by using a Karhunen-Loève transform to transform a feature vector into two-dimensional representations of the style of the authors, which determine a point in an image. It was found that the authorship assigned to these papers was consistent with that found in other studies, and that character n-grams were the best-performing feature type at that time. Another example is Juola (2008) in his work on Authorship Attribution, where he proposed the use of character n-grams and showed that they are particularly helpful for uncovering the writing style of authors.

Although character n-grams have been a useful feature in stylometry and authorship verification, there are several limitations and drawbacks to this approach. Firstly, as Kestemont (2014) points out, n-grams capture a wide range of information, including both style and content ("n-grams capture a bit of everything"). This can make it difficult to accurately identify and distinguish an author's writing style from the content of the text. To mitigate this problem, Stamatatos (2017) and Stamatatos (2018) proposed masking all words that are infrequent in a language. This approach, however, may not account for spelling mistakes or variations in writing commonly found on social media. Additionally, Koppel, Schler and Argamon (2009) have raised concerns about the caveats of character n-grams, specifically that many of them will be closely associated with particular content words and roots. This can lead to unreliable and inconsistent results, making it challenging to draw meaningful conclusions about an author's writing style.

2.3 DEEP LEARNING APPROACHES

Deep learning is a subfield of machine learning that uses neural networks to model complex patterns in data. In recent years, it has become increasingly popular in natural language processing tasks, including Authorship Verification and Attribution. The use of deep learning in Authorship Verification and Attribution is relatively new, but it has already shown promise in improving the performance of traditional methods. One of the key advantages of deep learning approaches is their ability to learn representations of the text that are more abstract and meaningful than traditional methods. This allows them to capture more subtle stylistic differences between authors, which can be useful for identifying authorship.

2.3.1 *Neural Networks*

Neural networks (NNs) are a type of machine learning algorithm that are inspired by the structure and function of the human brain. They consist of interconnected layers of artificial neurons that are designed to process and analyze complex data, such as images, text, and audio. NNs are capable of learning from data and can be trained to perform a variety of tasks, such as image recognition, natural language processing, and speech recognition. In recent years, the use of architectures based on neural network models has grown in popularity rapidly. This is mainly due to the advances in computational power, which have made it possible to train very large neural networks on massive amounts of data.

This rise in popularity can also be seen within the field of style modeling. NNs are particularly well-suited for capturing complex patterns in written texts due to their ability to learn high-dimensional representations of data and their capacity to handle large amounts of data. One of the earliest examples of NNs used in style modeling is the work of Kjell (1994). He explored the use of neural network classifiers for authorship recognition on the Federalist Papers by analyzing the relative frequencies of letter pairs within text samples. Although the classification of the twelve papers of uncertain authorship was inconsistent, the study already highlighted the potential of neural network classifiers for the task of Authorship Attribution. One notable success that fulfilled this potential was that of Koppel, Argamon and Shimoni (2002), where they trained a feedforward NN to distinguish between the writing styles of authors. They used a combination of words, punctuation, and capitalization as input features and achieved high accuracy in Authorship Attribution.

Other researchers have explored the use of NNs in style modeling by using different types of architectures, such as the use of Convolutional Neural Networks (CNNs). CNNs are a type of deep learning algorithm that are particularly well-suited for processing and analyzing sequential data, such as text. The effectiveness of this type of architecture is demonstrated in several studies, including Shrestha et al. (2017) and Ruder, Ghaffari and Breslin (2016). These studies have shown that CNNs are able to outperform the previously

mentioned methods for Authorship Attribution on large datasets. The ability of CNNs to process character-level information, in addition to word-level information, is one of the main reasons why they are able to achieve better performance. The ability to process character-level information allows CNNs to capture subtle stylistic patterns that might be missed by traditional methods that rely on word-level information only. Additionally, the use of convolutional layers in CNNs allows them to automatically learn and extract the most informative features from the input text, which further improves their performance.

Despite the success of neural networks in style modeling, there are still challenges and limitations that need to be addressed. One of the main issues is the need for large amounts of training data to achieve good performance. This can be a problem for low-resource languages or for authors with a small number of samples. Additionally, the interpretability of neural network models is often limited, making it difficult to understand the factors that contribute to their predictions. This can be especially problematic for style modeling, where understanding the underlying factors of the written style might be of interest. Furthermore, NNs and CNNs are highly complex models that require significant computational resources, making them less suitable for real-time applications. Finally, there is still room for improvement in terms of robustness and generalization, especially when dealing with noisy or out-of-domain data.

2.3.2 *Recurrent Neural Networks*

Recurrent Neural Networks (RNNs) are a type of neural network used to process sequential data, such as text. They have a memory component called the hidden state, which is updated at each time step to make predictions. RNNs can be unrolled to process a sequence of inputs, and the hidden state at each time step is used to make predictions. The most common types of RNNs are Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, introduced by Hochreiter and Schmidhuber (1997) and Cho et al. (2014), respectively.
Some studies have used RNNs for the Authorship Attribution and Verification tasks. Zhao et al. (2018), for example, used RNNs and an attention mechanism to capture important information in the sequence, showing that this approach was superior to traditional machine learning methods, such as Support Vector Machines (SVMs). Jafariakinabad, Tarnpradab and Hua (2019) introduced a syntactic recurrent neural network to encode the syntactic patterns of a document - learned from the sequence of part-of-speech (POS) tags - which outperformed lexical models for authorship attribution.
Encoder-decoder models, consisting of an encoder RNN and a decoder RNN, have been applied to natural language processing tasks such as machine translation, text summarization, and image captioning (Cho et al., 2014; Nallapati, Xiang and Zhou, 2016; Xiao et al., 2019). They work by encoding the input sequence into a context vector and then decoding it into the output sequence. Lample et al. (2019) used this architecture for style transfer in natural language, but this is not directly applicable to authorship

verification.

RNNs, while powerful in modeling sequential data, have several limitations when it comes to style modeling. One issue is the difficulty of capturing long-term dependencies in text, as the hidden state of an RNN can only capture information from a limited context. Additionally, RNNs struggle to handle the large amounts of data required for style modeling, as the computational complexity of training RNNs grows with the size of the input sequence. Furthermore, RNNs are sensitive to the ordering of the input, making it difficult to model the style of text written in different languages or from different domains.

2.3.3 *Transformer-based architectures*

Transformer-based architectures are a recent development in the field of natural language processing, first introduced by Vaswani et al. (2017). These architectures are based on the transformer, a novel mechanism for attention-based neural networks. The transformer allows for the parallel computation of attention mechanisms, significantly increasing the efficiency and effectiveness of the model. Before the transformer, RNNs were the dominant architecture for natural language processing tasks (see § 2.3.2). Convolutional neural networks were also used for some natural language processing tasks, but they struggle to model dependencies between different positions in the input sequence (see § 2.3.1). The transformer architecture addresses these issues by introducing self-attention mechanisms, which allow the model to weigh the importance of different positions in the input sequence. This allows the transformer to better capture long-term dependencies and global context, resulting in improved performance on a variety of natural language processing tasks.

Since its introduction, transformer-based architectures have been widely adopted and have achieved state-of-the-art results on a number of natural language processing tasks, including language translation, language modeling, and text summarization (Liu et al., 2020; Wang, Li and Smola, 2019; Liu and Lapata, 2019). Some examples of transformer-based models are: BERT (Devlin et al., 2019), GPT-4 (OpenAI, 2023), and RoBERTa (Liu et al., 2019).

Transformer-based architectures have been increasingly used in recent years for the Authorship Verification and Authorship Attribution tasks, as well as for style modeling. These models, such as BERT and its variations, have been shown to be powerful in extracting features from text, making them well-suited for these tasks. For example, Rivera-Soto et al. (2021) have used transformer-based architectures for domain transfer in the AV task. The researchers discovered that there was a substantial degree of transferability across the Reddit, Amazon reviews, and fanfiction domains that were tested. Additionally, they found that models trained on the Reddit dataset demonstrated consistently strong transfer performance. Wegmann, Schraagen and Nguyen (2022) introduced a variation of the AV task that controls for content using conversation or domain labels. They found

that representations trained by controlling for conversation are better at representing style independent from content than representations trained with domain or no content control. Manolache et al. (2021) have proposed five new public splits over the *PAN_2020* dataset[2], specifically designed to isolate and identify biases related to the text topic and to the author's writing style. They found that models trained without the named entities obtain better results and generalize better when tested on DarkReddit, a new dataset for AV. Furthermore, Zhu and Jurgens (2021) proposed a new approach to studying idiolects through a massive cross-author comparison to identify and encode stylistic features. The neural model achieves strong performance at authorship identification on short texts and through an analogy-based probing task, showing that the learned representations exhibit surprising regularities that encode qualitative and quantitative shifts of idiolectal styles. Fabien et al. (2020) presented a deep learning-based approach for Authorship Attribution that fine-tunes a pre-trained BERT language model and explicitly includes additional stylistic features. This approach achieves competitive performance on multiple datasets, outperforming state-of-the-art models at the time.

While transformer-based architectures have been shown to be powerful in the AV and AA tasks, they also have limitations. For example, works such as Fabien et al. (2020) and Rivera-Soto et al. (2021) lack control for content, meaning that the model may not be able to distinguish between style and content, which can lead to models that are biased towards certain topics or named entities. Additionally, while these studies utilize data from diverse sources like Reddit, they don't fully exploit the range of writing conventions and styles present across different social media platforms. For instance, Marko and Buker (2022) emphasized that each social media platform has its unique writing conventions and styles, suggesting the value of using varied and representative data sources for training models. While the value of leveraging data from diverse social media platforms to capture a broader range of writing styles seems promising, it's important to note that exploring this avenue is beyond the scope of this thesis.
Other research, such as that by Wegmann, Schraagen and Nguyen (2022) does include some level of content control, but this can still be improved upon, as is suggested in Chapter 3 of this thesis.
Another limitation of transformer-based architectures for style modeling is their lack of interpretability. Transformer models are highly complex and their internal workings are not easily understood, making it difficult to determine why a particular prediction was made. This can limit their usefulness for certain applications, such as in situations where the reasoning behind a prediction is crucial. However, it is important to note that this issue of interpretability is not the primary focus of this thesis.

2 https://pan.webis.de/clef20/pan20-web/author-identification.html

2.4 EVALUATION METHODS

Evaluating style representations poses several challenges, which must be considered when developing and comparing different approaches. Firstly, there is no agreed-upon definition of writing style, which makes it difficult to determine the criteria for a good style representation. Secondly, there is limited annotated data available for training and evaluating style representations, which can impact the performance and generalizability of models. Additionally, there is a lack of standard evaluation metrics for style representation, which makes it difficult to compare the performance of different models. Finally, style is subjective, and what one person considers a good representation of style may not align with another person's perspective. These challenges must be carefully considered when designing experiments to evaluate style representations, and appropriate methods must be used to mitigate their impact on the results.

The **STEL** (STyle EvaLuation) framework, proposed by Wegmann and Nguyen (2021) is a modular, fine-grained, and content-controlled similarity-based evaluation framework for testing a model's ability to capture the style of a sentence. The framework has been designed to test the style-measuring capability of different models, and it uses tasks that require ordering sentences to match the order of anchor sentences based on their style. An example of such a task is illustrated in Fig. 2.1.
The framework uses different characteristics, such as contraction and number substitution, and more general dimensions of style, such as formal/informal and simple/complex. By using both complex style dimensions and simpler characteristics, STEL allows for very controlled and fine-grained testing, meaning that it is able to test for small, subtle differences in style, rather than just broad distinctions. This allows researchers to easily make sure that only the characteristics and no other aspects change.

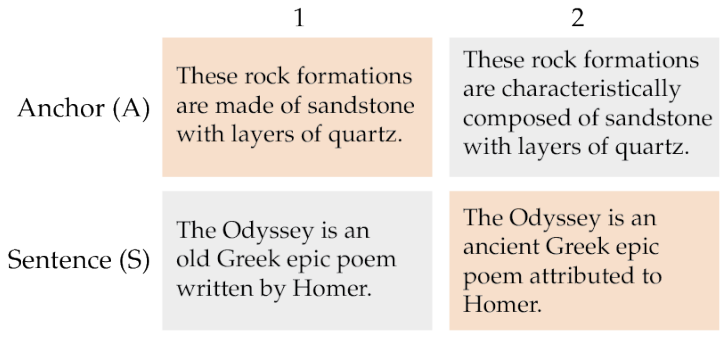


Figure 2.1: A STEL task instance. Anchor 1 (A1) and anchor 2 (A2) and the alternative sentences 1 (S1) and 2 (S2) are split along one of the proposed style dimensions: simple/complex. The task is to order S1 and S2 to match them to the same style as A1 and A2. In this figure, the utterances belonging to the same style have the same background color. Here, the correct order is thus S2-S1.

One of the main challenges in evaluating style is disentangling it from content. To control

for content, the framework uses parallel paraphrase datasets, which consist of a set of sentences written in one style and a parallel set of sentences written in another. This allows for a more accurate evaluation of a model's ability to capture style, independent of content. The STEL framework contains several components, including formal/informal and simple/complex dimensions, and contraction and number substitution characteristics. These components are designed to be easy to identify, allowing for a more straightforward evaluation of a model's style measuring capability.
As previously mentioned, one of the key innovations of the STEL framework is its ability to control for content. By using parallel paraphrase datasets, where sentences are written in different styles but convey the same content, the authors are able to control for content and more accurately evaluate the model's ability to capture style.

The STEL framework is evaluated using two different task setups, the quadruple and triple setups, which are similar to the triple and quadruple training instances in the field of metric learning. In the quadruple setup, the model is given four sentences, two anchors, and two test sentences, and is asked to order the test sentences to match the style order of the anchors. In the triple setup, the model is given three sentences, one anchor, and two test sentences, and is asked to decide which of the two test sentences matches the style of the anchor the most.

3

METHODS

This chapter outlines the steps taken to conduct the research presented in this thesis. The aim of this research is to investigate the use of transformer-based approaches for style representation, with an emphasis on controlling for content. In Section § 3.1, the data handling process is outlined, including the collection and pre-processing of data from Reddit. In Section § 3.2, the training task for the Authorship Verification task is described in detail, including the contrastive learning approach and the supervised Contrastive Authorship Verification task. The preparation for the input data and the fine-tuning process are also explained in that section. The evaluation of the proposed approach is described in Section § 3.3, including the use of a baseline model, the STEL framework, and performance metrics.

3.1 DATASET

In order to effectively train a transformer-based model for the Authorship Verification task, a substantial amount of data is required. Transformers, being a complex architecture, require an enormous amount of data to learn from in order to accurately capture the nuances of writing styles. For this thesis, I collected a dataset of 5,255,875 utterances, to ensure that my model has enough data to learn from. While in the field of Natural Language Processing, 5 million utterances may not seem excessive, it is a considerable amount in comparison to related work that used transformers for style representations.

A concise comparison with other studies, as demonstrated in Table 3.1, brings this amount of data into perspective. It illustrates the number of utterances and distinct authors utilized in their respective datasets, categorized by data sources. Of note is the variance in dataset sizes, from the relatively modest collection of Wegmann, Schraagen and Nguyen (2022) to the extensive corpus employed by Rivera-Soto et al. (2021).
It is important to mention that the dataset for this thesis, as represented in the table, is prior to the implementation of any pre-processing or selection procedures. Post-processing, the final quantity of data used is significantly reduced, with the specifics outlined in the subsequent subsection.

Given this landscape, the choice of a 5 million utterance dataset in this thesis was motivated by several considerations.

Paper	Total Utterances	Total Authors	Data Source	Utterances	Authors
Wegmann, Schraagen and Nguyen (2022)	630,000	385,157	Reddit	630,000	385,157
Zhu and Jurgens (2021)	1,843,130	184,313	Reddit	553,680	55,368
			Amazon reviews	1,289,450	128,945
Rivera-Soto et al. (2021)	113,778,169	1,176,000	Reddit	100,000,000	1,000,000
			Amazon reviews	13,500,000	135,000
			Fanfiction	278,169	41,000
This thesis	5,255,875	1,038,249	Reddit	5,255,875	1,038,249

Table 3.1: Comparison of utterances and authors used by relevant related works, broken down by data source. In the case of this thesis, the listed figures represent data quantities before the implementation of any pre-processing or selection steps. The final quantity of data utilized, post-processing, is notably less and will be detailed in the ensuing section.

Firstly, it is important to note that, in this work, I am not training the model from scratch but I am fine-tuning an existing model. As a result, the model has already mostly only learned semantic embeddings from a large-scale dataset. Since the goal is to shift the model's focus from semantic to stylistic embeddings, a substantial amount of data is still necessary for effective fine-tuning.
Secondly, a balance between dataset size and available computational resources was essential. Although Rivera-Soto et al. (2021)'s datasets are considerably larger than the dataset used in this work, the chosen dataset size in this thesis is still sufficient for the task at hand, given the dataset size and performance of the model from Wegmann, Schraagen and Nguyen (2022).
The chosen dataset size for this thesis offers a promising foundation for the investigation of the model's performance on stylistic tasks, particularly given that I am fine-tuning the model rather than training it from scratch. Yet, it is important to bear in mind that there is an opportunity for further exploration in this area. As a suggestion for future work, one could conduct experiments with larger datasets. This would provide insights into whether an increase in data volume could potentially enhance the model's performance, and it would be interesting to compare these findings with the outcomes of the current approach.

The data for this study was collected from Reddit. The reason for using social media data as opposed to other types of data is that social media platforms have become a prevalent form of communication for a great number of people and have a vast amount of written content that is publicly accessible. Additionally, social media platforms have a large user base, which provides a wide range of writing styles to learn from.
In this subsection, I describe the process of collecting my data. An overview of the number of utterances and authors can be found in Table 3.3.

COLLECTION The data was collected using the *ConvoKit* tool developed by Chang et al. (2020). This tool allows for the collection of conversation threads and comments from a variety of subreddits, providing a large and diverse sample of written language. Specifically, I used the Reddit Corpus generated by Wegmann, Schraagen and Nguyen (2022). This

is a sample of conversations (i.e., comment threads) on Reddit from 100 active subreddits[1] from 2018. For each subreddit, there are at least 600 comment threads, each having at least 10 comments. This leaves a total of 5,255,875 utterances by 1,038,249 unique authors from a total of 60,000 conversations.

PRE-PROCESSING To prepare the data for further use, I applied several pre-processing steps. A full overview of the number of removed authors and utterances for each step can be found in Table 3.2. Steps that do not remove any utterances or authors were left out for brevity.

Firstly, the data was cleaned by masking irrelevant information and text that can not be attributed to style. This includes user mentions - indicated by the "/u/" or "u/" prefix - as well as subreddit mentions — indicated by the "/r/" or "r/" prefix. These mentions were masked by replacing them with the "[MENTION]" token.
Secondly, all URLs were masked by replacing them with the "[URL]" mask. Care was taken that stylistic features from these URLs were not removed during this process. Given that Reddit utilizes Markdown for links, the process involved removing only the "link" part, not the text itself. This way the stylistic difference between formatted hyperlink text (e.g., "Take a look at Utrecht University!"), and a regular link (e.g., "Take a look at Utrecht University: https://www.uu.nl/!") was retained.
The third pre-processing step is specific to Reddit and includes the removal of utterances that contain the "RemindMe!" bot command. This bot lets you set a reminder but does not actually contribute anything to a discussion.
Fourthly, all other invalid utterances were removed. All utterances of only spaces, tabs, line breaks, emojis, or of the form: "[deleted]", "[deleted]", "[removed]", and "[removed]" were disregarded. Utterances that contained only mentions, URLs, or a combination of both were also removed.
In the fifth step, all utterances by invalid authors or authors for whom the unique authorship of the utterances could not be verified were removed. This includes authors with the username "[deleted]", but also the "AutoModerator" bot, as well as any user who either has "bot" in their username or uses the word "bot" in all of their utterances. While it is possible that some non-bot authors may have been filtered out, I expect this number to be non-significant, as this is only 1.51% of the total authors in the dataset.
The sixth step consisted of removing all authors who only had one utterance in the dataset. The reason for this removal is that it is not possible to find other utterances written by the same author.
Finally, in the seventh step, all utterances that were too long to fit in the RoBERTa model ($>$ 512 tokens) were removed from the dataset. This was done by using the RoBERTa tokenizer to tokenize each utterance and filtering the utterances based on their length.

1 https://zissou.infosci.cornell.edu/convokit/datasets/subreddit-corpus/subreddits_small_sample.txt

Preprocessing step	Utterances removed	Authors removed	% Utterances removed	% Authors removed	Remaining utterances	Remaining authors	Utterances / author
RemindMe!	1,526	344	0.03%	0.03%	5,254,349	1,037,905	5.062
Invalid utterances	488,979	5,109	9.31%	0.49%	4,765,370	1,032,796	4.614
Mention/URLs	25,351	4,753	0.53%	0.46%	4,740,019	1,028,043	4.611
Invalid authors	51,919	3	1.10%	0.00%	4,688,100	1,028,040	4.560
Likely bots	164,599	15,488	3.51%	1.51%	4,523,501	1,012,552	4.467
1 utterance	474,339	474,339	10.49%	46.85%	4,049,162	538,213	7.523
Too long	10,415	14	0.26%	0.00%	4,038,747	538,199	7.504

Table 3.2: Summary of preprocessing steps and their impact on the dataset, showing the number of removed utterances and authors, percentage of data removed, and remaining data after each step, along with the ratio of remaining utterances per author. Steps that do not remove any utterances or authors were left out for brevity.

DATA SUMMARY After the pre-processing, 4,038,747 utterances from 538,199 unique authors remained. That means that there is an average of 7.5 utterances per author, and the maximum number of utterances by one author is 2,962. These utterances were sampled from a total of 59,962 conversations. This means that the data contains almost all conversations from the original corpus. As suggested by Wegmann, Schraagen and Nguyen (2022), preserving such a wide array of conversations potentially enhances the generalization of style embeddings, as it offers a more comprehensive representation of stylistic diversity and variability across different conversational contexts.

Before pre-processing			After pre-processing		
# Utterances	# Authors	Avg. per author	# Utterances	# Authors	Avg. per author
5,255,875	1,038,239	5.06	4,038,747	538,199	7.50

Table 3.3: Overview of the number of utterances and number of authors for the dataset before and after pre-processing.

3.2 TRAINING TASK

Recall that in this thesis, the primary goal is to enhance the representation of writing styles by building upon and refining transformer-based approaches. To achieve this, I focused on the Authorship Verification (AV) task, which is a binary classification problem that aims to determine whether two given texts are written by the same author. Specifically, I experimented with controlled content during the fine-tuning process by introducing semantically similar texts, which should encourage the model to concentrate on stylistic nuances rather than content. By training models on this task, the models should not only learn to distinguish between authors by the difference in topics but also by nuances in their writing style and thus - hopefully - encode writing style in the learned representations. The traditional approach to AV is to train a model on a labeled dataset of text pairs, where each pair is labeled as either "same author" or "different author", following a supervised learning paradigm. In recent years, contrastive learning has emerged as an approach that can enhance the representation of natural language in machine learning

models. Contrastive learning involves learning representations by comparing similar and dissimilar examples. Essentially, an anchor example is chosen, and positive (i.e. written by the same author as the anchor) and negative (i.e. written by a different author than the anchor) examples are identified with respect to this anchor. The model then learns to bring the anchor and positive examples closer in the representation space, while pushing the anchor and negative examples apart. This method is not purely prediction-based, which sets it apart from some traditional supervised and unsupervised learning methods. It's also worth noting that this approach leverages the structure within the data itself to learn meaningful representations, much like some other machine learning techniques. Examples of a training instance of both a positive pair and a negative pair can be found in Fig. 3.1.

Figure 3.1: Two example text pairs for the AV task (these examples are made up). The task is to determine whether Utterances 1 (U1) and 2 (U2) were written by the same person or not. In the case of the first example, both utterances were written by the same author. This is thus a positive example. In the case of the second example, both utterances were written by a different author. This is thus a negative example.

In this thesis, I applied the supervised contrastive learning approach as proposed in "SimCSE: Simple contrastive learning of sentence embeddings" (Gao, Yao and Chen, 2021), which was designed to learn semantic similarity representations, to the AV task. While the method has shown promising results in capturing semantic information, it might not yield the same results when applied to learning stylistic representations. In the SimCSE paper, a contrastive loss function is utilized, an idea originally conceived by Hadsell, Chopra and LeCun (2006) for binary classification tasks. This concept of contrastive learning involves bringing closer the representations of similar instances and distancing dissimilar ones. In my work, I adopt the same contrastive loss function as employed in the SimCSE paper, thus following the path that Hadsell, Chopra and LeCun (2006) blazed. The specific form of the loss function can be found in Equation 3.1. This type of loss function is particularly effective for learning representations of text, as it allows the model to learn to distinguish between similar inputs (i.e. in this case, the writing style of different authors). This is important for this research, as the goal is to be able to accurately distinguish the writing style of individual authors in order to learn style representations. The contrastive loss function was used to fine-tune the model on the collected data for the AV task. In the paper by Gao, Yao and Chen (2021) this has led to state-of-the-art semantic sentence

embeddings, and the approach is expected to result in a better representation of writing styles that can be used for the AV task, as is shown in both Zhu and Jurgens (2021) and Wegmann, Schraagen and Nguyen (2022). While the latter two papers do not use the exact same loss function as proposed by Gao, Yao and Chen (2021), they both employ contrastive learning approaches to achieve improved performance in capturing writing styles for the AV task.

$$l_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} (e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau} + e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^-)/\tau})}$$

Equation 3.1: The training objective for a set of paired examples $\mathcal{D} = (x_i, x_i^+, x_i^-)_{i=1}^m$ is shown above, where:
- m represents the total number of paired examples,
- x_i is the anchor utterance,
- x_i^+ is another utterance by the same author (positive example),
- x_i^- is an utterance by a different author (negative example),
- $\mathbf{h}_i$, $\mathbf{h}_i^+$, and $\mathbf{h}_i^-$ denote the representations of x_i, x_i^+, and x_i^-, respectively,
- τ is the temperature hyperparameter,
- *sim* is the cosine similarity function,
- N is the mini-batch size.

This is a batched contrastive loss function, aiming to minimize the similarity between the anchor example and all other negative examples within the same batch while maximizing the similarity between the anchor and its paired positive example. The final loss for a batch is the average of the l_i for all i in the batch.

The aforementioned loss function is implemented in the *Sentence-Transformers* library as the *MultipleNegativesRankingLoss* (Reimers and Gurevych, 2019). This implementation takes other inputs from the same batch into account, meaning that the model learns to distinguish between different positive and negative pairs in the same batch. The loss function behaves in such a way that it encourages the model to produce higher similarity scores for positive pairs (anchor and positive examples) and lower similarity scores for negative pairs (anchor and negative examples). The loss is minimized when the similarity score between the anchor and the positive example is higher than the similarity scores between the anchor and all the negative examples in the batch. By learning to rank multiple negatives, the model becomes better at distinguishing between the writing styles of different authors. This approach is different from other loss functions in the library as it explicitly focuses on ranking and comparing multiple negative examples. It has been shown to improve the performance of the semantic representations compared to using only one positive and one negative example per batch (Wang et al., 2020).

To apply the supervised contrastive learning approach to the Authorship Verification task and create the Contrastive Authorship Verification (CAV) task, three more steps need to be taken. First, I need to mine possible paraphrases (§ 3.2.1). This is done to generate a rich set of semantically similar examples, which serves as the backbone of the training task. By doing so, I can create challenging negative examples that are integral to my approach of improving style representations. After that, I need to pair the texts,

such that the input examples for the model consist of (Anchor, Positive, Negative) triplets (§ 3.2.2). Finally, I need to fine-tune the transformer itself (§ 3.2.3).

3.2.1 *Paraphrase mining*

First, it is important to select semantically similar pairs of utterances for training the model. The goal of this approach is not only to predict whether a text pair is a "positive" or "negative" example, but it also aims to develop a nuanced understanding of the texts by learning style-specific representations that capture the distinctive characteristics of each author's writing. By including text pairs that are semantically similar, the model should focus on stylistic variations rather than content-based differences, as the content itself is controlled for and remains relatively constant across pairs. This way, the model is encouraged to learn features that are more specific to an author's writing style and less dependent on the content of the texts.
To illustrate the importance of content control, consider an example where we do not control for content. Suppose we have two text pairs for training: one pair where two texts discuss the same topic, "climate change," for example, and another pair where the texts discuss completely different topics, one being "climate change" and the other a "music concert." If the model isn't guided to focus on stylistic variations, it might learn to differentiate authors based on the content of their texts rather than their unique writing styles. In this case, the model could incorrectly attribute the differences between the texts discussing "climate change" and "a music concert" to stylistic variations rather than the stark difference in their content.

To find these paraphrases, I used the *paraphrase_mining* utility function from the *Sentence-Transformers*[2] library (Reimers and Gurevych, 2019). The first step in this process involves creating semantic embeddings for each text. In this thesis, I used the "all-mpnet-base-v2" model[3], since this model is considered the best quality general-purpose model offered by the library and is suitable for generating high-quality semantic embeddings. Once these semantic embeddings are generated, they are compared using the cosine similarity function. This function finds and ranks the *top-k* semantically similar sentences for each sentence in the dataset. For this thesis, the *top-k* parameter was set to 100, meaning that for each sentence, the function attempts to identify up to 100 other sentences as potential paraphrases. The *max_pairs* parameter was set at 100,000,000. This parameter indicates the maximum number of paraphrase pairs that the function returns across the entire dataset, not just for one utterance. It was set to a large number to ensure that I retrieve as many paraphrase pairs as possible across all utterances in the dataset. It's important to note that the *top-k* and *max_pairs* parameters serve different purposes. While *top-k* limits the number of paraphrase candidates for each individual sentence, *max_pairs* limits the total number of paraphrase pairs that are returned from the entire dataset. It

2 https://www.sbert.net/examples/applications/paraphrase-mining/README.html
3 https://huggingface.co/sentence-transformers/all-mpnet-base-v2

should be noted that I do not set a range on semantic similarity, but it is instead limited by taking the best 100,000,000 paraphrases.

In order to ensure the quality of the paraphrase pairs, I temporarily hid two types of texts from utterances before the mining. More specifically, I first removed all lines from a text that start with the ">" symbol, since this indicates a quoted text, and often implied a reply to another text in the same conversation. This would result in an unfairly high similarity score since both texts have a lot of overlap of the utterance. Secondly, I also removed texts that only consisted of emojis because the paraphrase mining model cannot deal with such texts. If a sentence consisted only of emojis, it would be considered a paraphrase of every other sentence that also consisted only of emojis, resulting in a similarity score close to 1, and a lot of "similar" paraphrases.
After mining the paraphrases, I also performed a filtering step to remove sentences by the same author, and duplicate sentences, where sentence 1 and sentence 2 of a paraphrase are the same. This was done to ensure the quality of the resulting paraphrase pairs.

The result of this process is a set of text pairs, where the first text is semantically similar but written by a different author than the second text. An example of a text pair that is semantically similar but written by a different author is illustrated in Fig. 3.2.

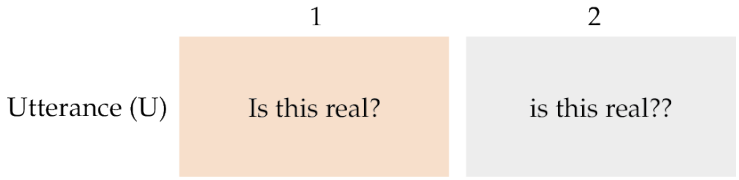


Figure 3.2: An example text pair for the AV task where the texts are semantically similar (cosine similarity of 0.93) but stylistically different. In this case, both utterances were written by a different author. This is thus a negative example.

3.2.2 *Pairing utterances*

The main novelty of the approach presented in this thesis lies within the combination of the previous step (§ 3.2.1) and this one. In this step, I paired each utterance from an author with positive examples (i.e., other utterances from the same author) and negative examples (i.e., utterances from different authors). For the positive examples, I paired each utterance with all other utterances written by the same author. For the negative examples, I selected the semantically similar utterances that are described in the previous subsection. Despite the intent to match as many utterances as possible with semantically similar counterparts, not all data could be paired this way. Therefore, for the remaining data, I generated the negative pairs by uniformly sampling texts written by a different author. This was done to ensure the model also gains experience dealing with dissimilar pairs and doesn't rely solely on semantic similarity for Authorship Verification.

The first phase of my iterative experiment focused on manipulating the percentage of paraphrases in the training data. Specifically, I explored how changing this percentage affects the performance of the RoBERTa models. The aim was to determine whether increasing or decreasing the concentration of semantically similar negative examples influences the model's ability to distinguish writing styles. After gathering insights from this initial experiment, I conducted subsequent experiments, each building on the findings of the previous. This iterative process was designed to continually refine and optimize the performance of the RoBERTa models.

3.2.3 *Fine-tuning pre-trained transformers*

The fine-tuning process involved training the model on the collected data, to improve its ability to distinguish and verify the writing style of individual authors. To achieve this, I used the *Sentence-Transformers* library (Reimers and Gurevych, 2019) to fine-tune several "roberta-base" models (Liu et al., 2019). This model is a transformer-based neural network language model that has been trained on a large corpus of text data, making it effective for a range of natural language processing tasks, including Authorship Verification (Zhu and Jurgens, 2021; Wegmann and Nguyen, 2021; Wegmann, Schraagen and Nguyen, 2022). I experimented with different values for hyperparameters such as the loss function, learning rate, batch size, and the number of epochs during fine-tuning. In the case of the *MultipleNegativesRankingLoss*, the batch size is more important since it compares each example with other examples in the same batch. Having a larger batch size thus allows the model to compare to more negative examples.

3.3 EVALUATION

3.3.1 *STEL framework*

In this study, I use the STyle EvaLuation (STEL) framework as proposed by Wegmann and Nguyen (2021) to evaluate the performance of the models (see Section **??**). In addition to the standard STEL framework, I also employed the STEL-or-content task introduced by Wegmann, Schraagen and Nguyen (2022). This task presents an additional challenge as it requires the model to differentiate between writing styles and content. In this task, the model has to choose between two options: one that matches the anchor style but with unrelated content, and another that matches the content but with a different writing style. A visual representation of the STEL-or-content task is provided in Fig. 3.3.

I used the framework to evaluate the model's performance on specific characteristics of style such as contraction and number substitution. I also utilized the framework to evaluate the model's performance on more general dimensions of style like formal/informal and simple/complex. I used the STEL evaluation results and the STEL-or-content task

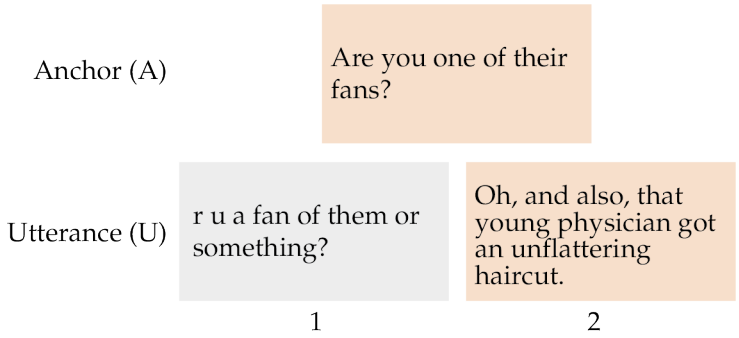


Figure 3.3: An example instance of a STEL-or-content task on the formal/informal dimension. In this case, the formal style of the Anchor sentence matches the style of Utterance 2, even though the content overlaps more with Utterance 1.

to determine whether the model has improved over previous work, and how much the addition of semantically similar text pairs has impacted the final results.

3.3.2 *Authorship Verification performance*

In addition to the evaluation using the STEL framework to determine how much the model controls for content, I also evaluated my model on the Authorship Verification and Contrastive Authorship Verification tasks. These evaluations primarily utilize accuracy as the performance measure, a choice motivated by the balanced nature of my data. While other metrics such as precision, recall, or F1 score could provide additional insight, especially in cases of class imbalance or when false positives and false negatives have significantly different costs, I deemed accuracy to be sufficient for this study given the nature of the task and the data distribution. As such, accuracy can give a clear, easy-to-understand assessment of how well my models are performing on the AV and CAV tasks.
Recall that in the AV task, we input a pair of texts and predict if they share the same author, resulting in a binary output. The CAV task, conversely, inputs a triplet of texts and predicts which text matches the anchor in authorship.

For this evaluation, I created two test sets, each consisting of 100,000 input examples. The data was split into training, validation, and test sets using an 80:10:10 ratio. Importantly, I ensured that there was no overlap of authors between these splits. This non-overlapping split is crucial to ensure that the model is not exposed to any data it has seen during training when it is validated and tested, which would artificially inflate performance metrics.
One test set contains 100% semantically similar negative examples, while the other consists of 0% semantically similar (randomly sampled) negative examples. It is important to note that although these test sets have a big overlap in authors, they still differ somewhat. The reason for using both test sets is to thoroughly assess the performance and robustness of the models under different conditions. By evaluating the models on test sets with varying degrees of semantic similarity between the anchor and negative examples, I can gain

insights into how well the models generalize to different types of input data.

The test set with 100% semantically similar negative examples provides a more challenging evaluation environment, as the models need to focus on capturing the nuanced differences in writing styles rather than relying on content differences. This test set allows me to evaluate how well the models have learned to disentangle writing style from content during training.
On the other hand, the test set with 0% semantically similar (randomly sampled) negative examples represents a more traditional AV and CAV evaluation setup. This test set allows me to compare the performance of the models to existing research and baselines, as well as assess their ability to adapt to a wider variety of negative examples that may not necessarily be semantically similar to the anchor sentences.

Since I am working with *Sentence-Transformers* (Reimers and Gurevych, 2019), which generate embeddings rather than performing classification directly, I determine whether two texts are written by the same author using a threshold-based approach. First, I generate embeddings for the sentences in the validation set and calculate the cosine similarity for each pair. I then use these cosine similarity scores to calculate the Receiver Operating Characteristic (ROC) curve and determine the optimal threshold as the value that maximizes the difference between the True Positive Rate (TPR) and the False Positive Rate (FPR). In the test set, if the cosine similarity between two authors is greater than this threshold, I classify the pair as being written by the same author. If the cosine similarity is lower, I classify them as being written by different authors.

3.3.3 *Baseline*

When comparing the performance of the proposed model in this thesis, it is important to consider the performance of previous models that have been used for the Authorship Verification task. In this thesis, I compared the performance against one principal baseline in the field of Authorship Verification. This baseline is strong and well-established, providing a rigorous point of comparison.

The baseline chosen is from the work of Wegmann, Schraagen and Nguyen (2022). They introduced a variant of the Authorship Verification task that implements content control using conversation (same comment thread) or domain (same subreddit) labels. Their findings suggest that representations trained by controlling for conversation yield superior capabilities in representing style independently from content. Their model is publicly available on the Huggingface website[4], negating the need for retraining.

The performance of the proposed model in this thesis is evaluated against this baseline to gauge its effectiveness and its ability to generalize in representing writing styles for

4 https://huggingface.co/AnnaWegmann/Style-Embedding

the Authorship Verification task. This comparison will provide valuable insights into the model's proficiency in disentangling style from content and its potential for practical application in the broader field of style modeling.

4

EVALUATION

In this chapter, I present the evaluation of my model for writing style representation using the fine-tuned RoBERTa model and the *Sentence-Transformers* Python library. My approach involves an iterative process of refining the model to better capture stylistic differences between authors. I employ the constrastive version of the Authorship Verification (AV) task - the Contrastive Authorship Verification (CAV) task - as the training task to learn the representations. Although these tasks are valuable for training purposes, they may not fully reflect the nuances of style modeling, as they primarily focus on author differentiation. Consequently, I put less emphasis on the performance of the AV and CAV tasks during the evaluation of the model and focus more on the model's ability to capture stylistic features and its applicability to a broader range of style-related tasks.
The AV and CAV tasks are both split into two test sets: one where all the different-author examples are a paraphrase of the anchor - the 100% AV/CAV task - and one where all the different-author examples are chosen at random — the 0% AV/CAV task.

To evaluate the model's primary focus on style modeling, I use the STEL framework, and I employ the STEL-or-content evaluation. Both of these evaluation methods are further elaborated upon in Section § 3.3.

Before delving into the evaluation, I will initially analyze the quality of the paraphrases in § 4.1. This analysis will provide a deeper understanding of the characteristics of the paraphrases and set the groundwork for subsequent evaluations.

In Section § 4.2 I provide a comprehensive overview of the results of this research. After that, each subsequent subsection will discuss the methods, results, and implications for each iteration, allowing readers to follow the development of my model and understand the rationale behind the changes I made. Finally, in Section § 4.9, I engage in a more overarching discussion of the results, synthesizing the findings from the various models explored throughout this thesis and analyzing them within a broader context to underscore their implications and potential applications for the field of style representation. By presenting my research in this manner, I provide a comprehensive view of the model's evolution and the impact of different modifications on its performance in style modeling.

4.1 SEMANTIC SIMILARITY ANALYSIS

Since the novelty of my approach lies in the effective use of semantically similar different-author examples, it is paramount to verify the quality of these examples. In this section, I will do exactly that. The following subsections will dissect the paraphrases that I use as semantically similar examples, beginning with a look at the overall score distribution.

It's important to clarify that the term "paraphrase" as used in this thesis does not strictly adhere to its conventional usage in related literature. Instead, it specifically refers to the method employed in this study for sampling different-author examples. This distinction is crucial for the interpretation of the results and discussions presented in this work.

4.1.1 *Score analysis*

An important step in analyzing the quality of paraphrases is examining the overall distribution of the paraphrase scores since this is the criterion that the paraphrases are selected on. Fig. 4.1 shows a histogram of paraphrase scores, which helps visualize the distribution of scores across the dataset.

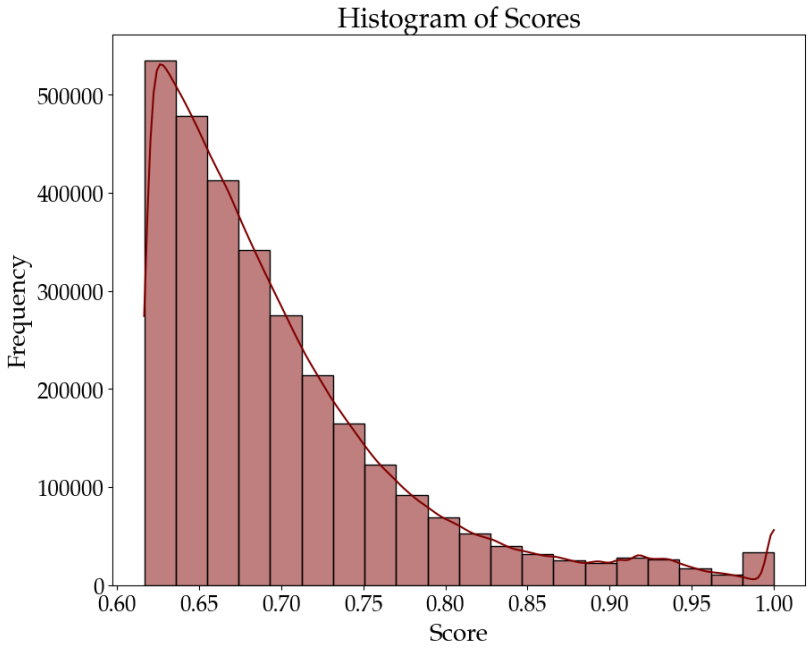


Figure 4.1: Histogram of paraphrase scores showing the distribution of scores across the dataset. The kernel density estimate (KDE) is also plotted to give a smooth estimate of the score distribution. The scores range from a minimum of 0.62 to a maximum of 1.00, with a median of 0.68. The 25th and 75th percentiles are 0.64 and 0.73, respectively.

As intuition would have us expect, this distribution is skewed towards a higher frequency of lower-scoring paraphrases. What also stands out, is the fact that the frequency of the scores in the last bin is higher than the bins just before it, breaking the trend of diminishing

frequencies.

If we look at Table 4.1, which shows 5 examples of the highest-scoring paraphrases, this disruption of the trend makes sense. Oftentimes, the sentences with similarity scores close to 1.0 differ only one or two characters. The lower the similarity scores, the more the anchor and paraphrases start to differ. Moreover, these top-scoring paraphrases predominantly consist of frequently used words or phrases in the English language. This observation is further explored and discussed in Section § 4.1.2 and the subsequent sections.

Anchor	**Paraphrase**	**Score**
And my Axe!	and My axe!	1.0
what an absolute unit	What an absolute unit	1.0
hot take	Hot Take	1.0
That's a great fight.	That's a great fight	0.976
AlAbAmA dOeSnT bElOnG iN tHe PlAyOfFs	AlAbAmA DoEs NoT DEseRvE To bE iN ThE PlAYoFfS	0.976

Table 4.1: Examples of the highest scoring paraphrases from the data. These examples represent paraphrases that are either identical or very similar to the anchor sentences, leading to high cosine similarity scores.

The fact that my dataset contains more lower-scoring paraphrases can also be explained by looking at Table 4.2. In that table, I listed 5 of the lowest-scoring paraphrase pairs that I sampled. Compared to Table 4.1, the paraphrases here differ a lot more semantically from the anchor. What is encouraging about this, however, is that even these lowest-scoring paraphrases still share at minimum the same topic, and usually the same sentiment about that topic as well.

While the tables Table 4.2 and Table 4.1 provide examples of the lowest and highest-scoring paraphrase pairs, respectively, they are not exhaustive representations of the entire dataset. To gain a more comprehensive understanding of the quality of the paraphrase pairs, I conducted a manual inspection of 100 randomly sampled paraphrases from across the score spectrum. This manual inspection confirmed that even the lower-scoring paraphrases generally maintained the same topic and sentiment as their respective anchors, despite their greater semantic differences. Therefore, based on this more extensive analysis, it can be inferred that the majority of the paraphrases in the dataset align with my goal of controlling for content. However, it's important to note that this conclusion is based on a sample and may not hold true for every single paraphrase in the dataset.

An interesting pattern that emerged from inspecting the dataset, and is illustrated by the examples in Table 4.1, is that not all paraphrases that differ only one or two characters get assigned the same similarity score. We can see that the difference between the anchor and the paraphrase in the fourth row is only the period at the end of the sentence, but the cosine similarity score it gets is not 1.0 (which would indicate full semantic equivalence). This indicates that the model used for the paraphrase mining is not always completely unaffected by differences in writing style. However, it's important to note that manual inspection of the data shows that these examples are representative of a broader pattern observed in the dataset, rather than isolated instances. A more comprehensive analysis

Anchor	Paraphrase	Score
Cute and inquisitive, really captures the intelligence of pigs in my opinion	What a cute pig! Is he/she just a small potbelly or?	0.6166
A medical condition? What are you, fucking Sanjay Gupta? Playing the goddamn music.	He says he thinks he may have a condition. But I'm not convinced he's a real doctor tbh.	0.6166
That looks cool, but at the time it wasn't actually used in any phones. When the iPhone was released, consumers had a choice between an actually decent capacitive screen, and shitty resistive screens that couldn't even properly register a light swipe. It's really no wonder that they picked the capacitive screen.	Oh yeah! That was such a cool concept too. It's such a bummer that nobody really takes risks with design or materials anymore, outside of the random phone here or there. Everything looks like a Samsung or the Essential/iPhone X.	0.6166
These birds are individuals who suffer just like cats or dogs and who fight for their lives just like cats or dogs. Consider that when you vote with your wallet and consider leaving them off your plate. This is coming from someone who has rescued dogs, cats, goats, chickens, guineas, sheep, etc. There are no differences between these animals in any way that truly matters.	Something saddens me when I see a bird as a pet!	0.6166
The Ravioli Lad, the Carbonara Kid.	Ravioli Ravioli give me the Formuoli	0.6166

Table 4.2: This table presents 5 of the lowest scoring anchor-paraphrase pairs that were used. These examples depict substantial variations in terms of length, wording, context, or semantic meaning compared to the anchor sentences, which result in low cosine similarity scores.

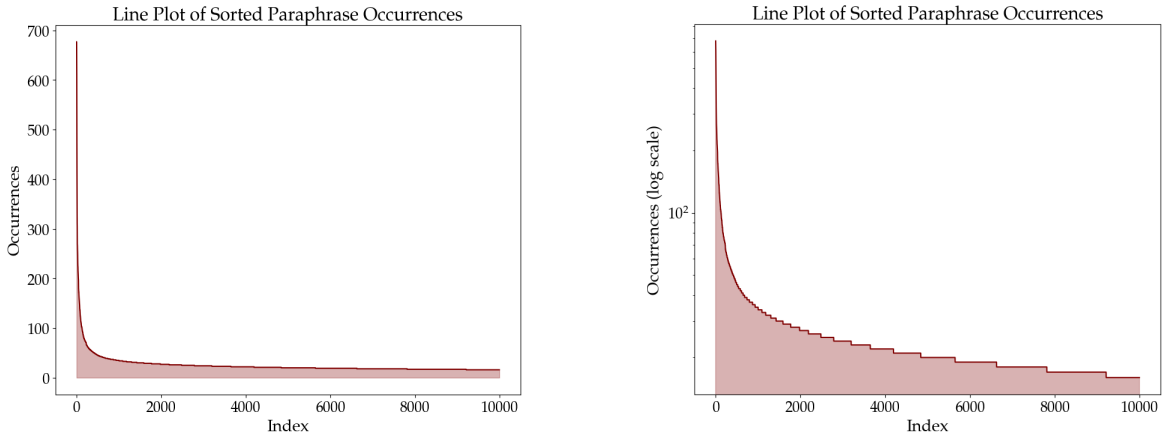
would be needed to determine the extent and impact of this potential sensitivity across the entire dataset.

4.1.2 *Frequency distribution*

Following the exploration of high and low-scoring paraphrases, I now move to examine the overall frequency distribution of all paraphrases in the dataset. When I refer to a paraphrase being "frequent in the dataset," it means that the same paraphrase appears multiple times. In the subsequent analyses, I only look at the mined paraphrases, and disregard their respective anchors.

To visualize this distribution, I created the plots presented in Fig. 4.2a and Fig. 4.2b. To create these plots, I first sorted the sampled paraphrases based on their occurrences, after which I plotted the occurrences of the first 10,000 indices (i.e., the 10,000 most frequently sampled paraphrases). I opted for only plotting the first 10,000 indices since the distribution is so heavily skewed towards the first indices that visualization on a larger scale would impair its explainability.

Fig. 4.2a and Fig. 4.2b show that this method of collecting paraphrases not only favors lower scoring paraphrases (as demonstrated in Section § 4.1.1) but is also skewed towards



(a) Line plot of paraphrase frequency distribution.

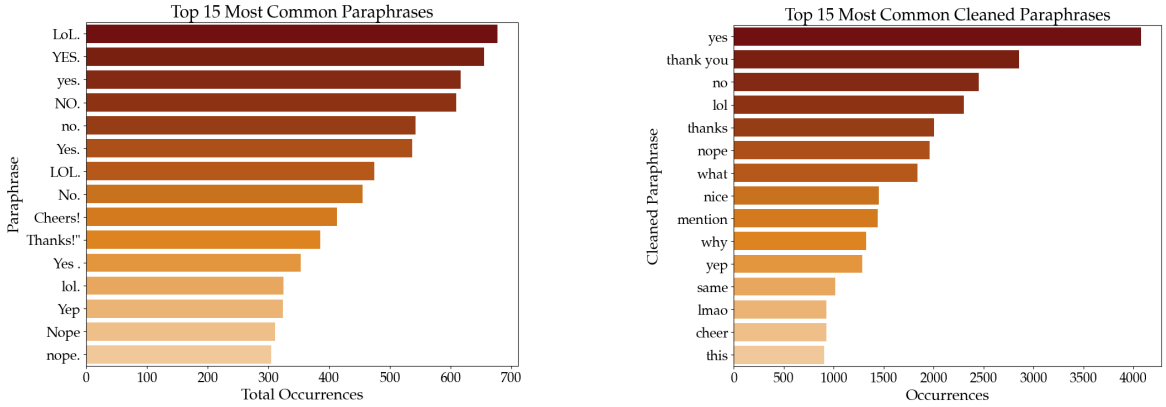
(b) Line plot of paraphrase frequency distribution in log scale.

Figure 4.2: Comparison of frequency distribution for the first 10,000 paraphrases in normal scale (left) and log scale (right), sorted by occurrences. The x-axis represents the index of the sorted paraphrases, while the y-axis represents the number of occurrences of each paraphrase. Both plots depict a rapid decrease in occurrence frequency, indicating a skewed distribution. The log scale on the right allows clearer visualization of the decline for less frequent paraphrases. The shaded area in both plots emphasizes the rate of decline.

reusing a small number of paraphrases a lot. To determine whether this apparent overrepresentation of a few paraphrases might be detrimental to the overall quality of the paraphrases, a closer inspection of these top paraphrases is necessary.
Assisting in this inspection, Fig. 4.3a shows the top 15 most common paraphrases. From an initial look-over of this graph, it seems that a lot of these paraphrases are very similar in meaning to some of the other paraphrases (e.g., "LoL." and "LOL."). This means that the diversity of these top paraphrases does not differ a lot semantically. This effect is especially illustrated in Fig. 4.3b, where I first applied some text cleaning - converting to lowercase, removing punctuation, and lemmatization - before plotting the most common paraphrases. Indeed, as depicted here, these similar utterances constitute a substantial proportion of the paraphrases. This is expected, given that short, one-word utterances such as "yes" and "thanks" are prevalent in the dataset (see Section § 4.1.3). The fact that Fig. 4.3a shows a lot of variations for the same paraphrases shows that there is a considerable amount of diverse writing styles for these utterances, which can be a positive factor for the training task. However, potential drawbacks could be that (i) these common paraphrases are disproportionately represented, and (ii) the dataset predominantly consists of shorter, similar paraphrases, while the occurrence of longer, unique utterances is relatively infrequent.

4.1.3 *Paraphrase word count and score*

The second concern highlighted in the previous section - the prevalence of shorter, similar paraphrases in the dataset - can be further elucidated by referring to Fig. 4.4. This figure, a hexbin plot, enables us to visualize the density and distribution of the data and also dis-



(a) Top 15 most common paraphrases.

(b) Top 15 most common paraphrases after cleaning.

Figure 4.3: Comparison of the top 15 most common paraphrases in the dataset before (left) and after (right) applying text cleaning processes such as converting to lower case, removing punctuation, and lemmatization. The total number of occurrences is represented on the x-axis in both plots.

cern potential correlations between the word count of the paraphrases and their respective scores. It should be noted that the color intensity of the hexagon bins corresponds to the log scale values.

Several key observations can be drawn from this plot. Firstly, paraphrases of shorter length appear to be favored across all scores. This is intuitive as text with less content is likely to have more semantically similar examples. Secondly, although they are still mostly shorter, paraphrases with lower scores tend to be more verbose. This can be attributed to the increased "freedom" in paraphrase selection associated with lower scores, which allows for more extensive expressions. This point is also illustrated in my interpretation of the lowest-scoring paraphrases in Table 4.2. The drawback of this phenomenon is that it may inadvertently lead to an overrepresentation of shorter, semantically similar phrases in the dataset. While these phrases indeed offer valuable insight into the nature of the most common conversational exchanges, their dominance may overshadow more complex, unique, and nuanced paraphrases which are essential for robust semantic understanding and variation in dialogues. Moreover, the trend of lower scores being associated with longer paraphrases may induce a bias in the model to perform better on simpler paraphrases over their complex counterparts, as the latter would likely be underrepresented in the training data.

4.1.4 *Comparison with Wegmann, Schraagen and Nguyen (2022)*

To highlight the differences between my approach for sampling negative examples and the method proposed by Wegmann, Schraagen and Nguyen (2022), I will give a short comparative analysis in this subsection. In their approach, for each utterance A1, another utterance B is randomly selected from the same conversation but written by a different author. Then, for all (A1, B)-pairs, a second utterance A2 is randomly chosen from all

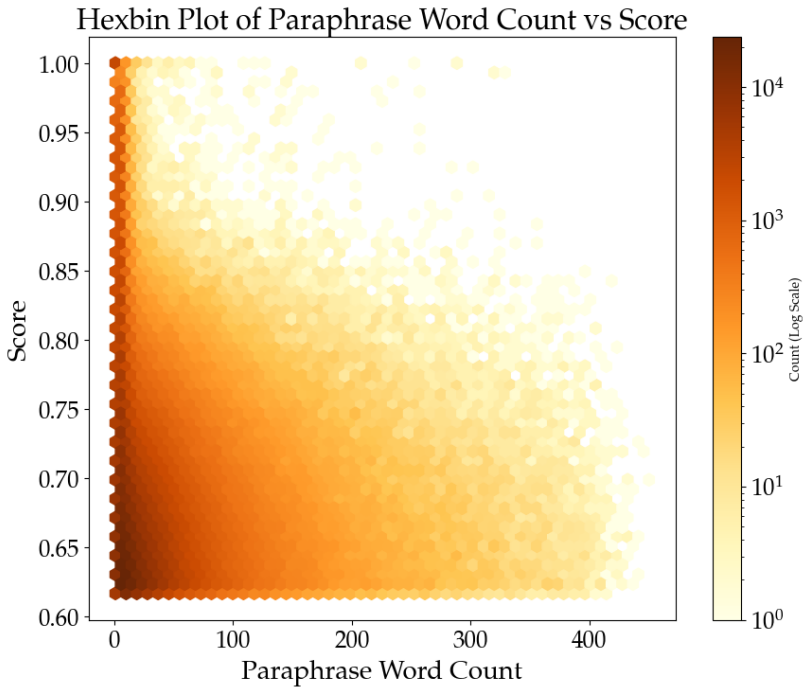


Figure 4.4: Hexbin plot comparing the word count of the paraphrases and their corresponding cosine similarity scores. The x-axis represents the word count in the paraphrases, and the y-axis represents the score associated with the paraphrases. The color of the hexagons represents the count of paraphrase-score pairs that fall into the area, with darker colors indicating higher counts, plotted on a log scale. This plot visualizes the density and distribution of the data, as well as any potential correlations between paraphrase word count and the score assigned by the model.

utterances written by the same author as A1, ensuring that A1 and A2 are not the same. Their version of content control is thus on the conversation level.

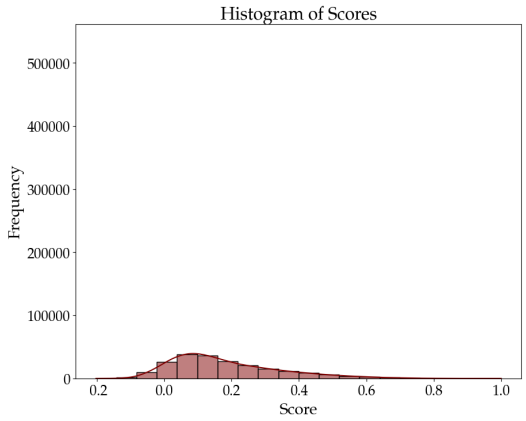
As the counterpart of Fig. 4.5b (or Fig. 4.1), Fig. 4.5a also presents a histogram of cosine similarity score frequencies. These scores were extracted from the training dataset of Wegmann, Schraagen and Nguyen (2022), using the same "all-mpnet-base-v2" model[1]. As expected, the distribution of scores is substantially different from that of my approach. The scores range from a minimum of -0.20 to a maximum of 1.00, with a median of 0.15. The 25th and 75th percentiles are 0.06 and 0.27, respectively. This contrasts sharply with the score distribution from my approach, which ranges from a minimum of 0.62 to a maximum of 1.00, with a median of 0.68 and 25th and 75th percentiles of 0.64 and 0.73, respectively.

The lower scores in the dataset of Wegmann, Schraagen and Nguyen (2022) are not surprising, given that their sampling approach is not based on similarity. Instead, it involves randomly selecting utterances from the same conversation but written by different authors, which can lead to a wider range of semantic differences and thus lower similarity scores.

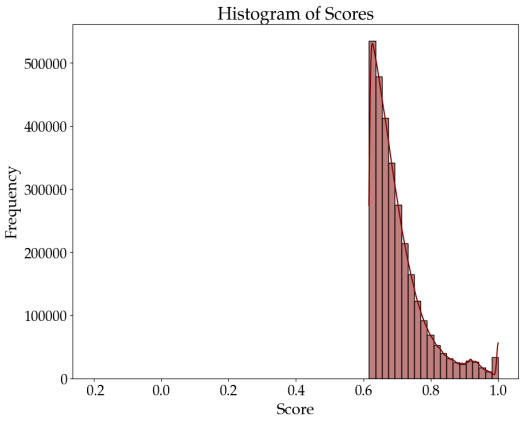
In contrast, my approach is similarity-based, which naturally leads to higher similarity

1 https://huggingface.co/sentence-transformers/all-mpnet-base-v2

scores. This difference in score distributions underscores the fundamental differences between the two sampling approaches and their potential implications for the resulting datasets.



(a) Histogram of different-author example cosine similarity scores across the dataset used by Wegmann, Schraagen and Nguyen (2022). The scores range from a minimum of -0.20 to a maximum of 1.00, with a median of 0.15. The 25th and 75th percentiles are 0.06 and 0.27, respectively.

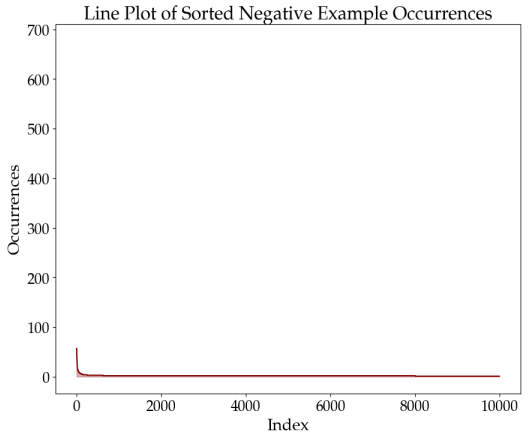


(b) Histogram of different-author example cosine similarity scores across the dataset used by me. The scores range from a minimum of 0.62 to a maximum of 1.00, with a median of 0.68. The 25th and 75th percentiles are 0.64 and 0.73, respectively.

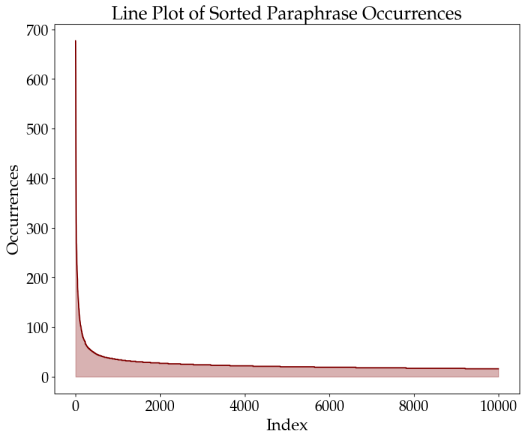
Figure 4.5: Side-by-side comparison of the distribution of cosine similarity of the anchor-negative example pairs. The left plot shows the histogram for the sampling method that Wegmann, Schraagen and Nguyen (2022) used, while my sampling method is highlighted in the plot on the right. The kernel density estimate (KDE) is also plotted to give a smooth estimate of the score distribution. For a fair comparison, both plots have the same x and y limits.

This difference in sampling approaches becomes even more apparent when examining the frequency distribution of the first 10,000 different-author examples, sorted by occurrences, as shown in Fig. 4.6a. The line representing the frequency distribution in their dataset "falls flat" to about 3 occurrences almost immediately. This suggests that there is a high level of diversity in their different-author examples, with most of them appearing only a few times in the dataset. In stark contrast, the frequency distribution of my approach, illustrated on the same axes in Fig. 4.6b, does not exhibit the same rapid decline. Instead, there is a more gradual decrease in the frequency of occurrences, indicating that the same paraphrases are sampled more frequently in my dataset. My similarity-based approach thus tends to sample the same paraphrases more frequently.

The distinctions between the two approaches are further highlighted in a hexbin plot comparing the word count of the different-author examples and their corresponding cosine similarity scores, as shown in Fig. 4.7a. Two key observations can be made from this plot. First, the word counts in their dataset are generally much higher than in my dataset (Fig. 4.7b). This discrepancy is likely due to two main factors: (i) their method of handling longer examples, where they opt to trim or "truncate" examples that exceed a certain length limit, whereas in my approach, I filter out and exclude any examples that



(a) Frequency distribution for the first 10,000 different-author examples used by Wegmann, Schraagen and Nguyen (2022), sorted by occurrences.



(b) Frequency distribution for the first 10,000 different-author examples used by me, sorted by occurrences.

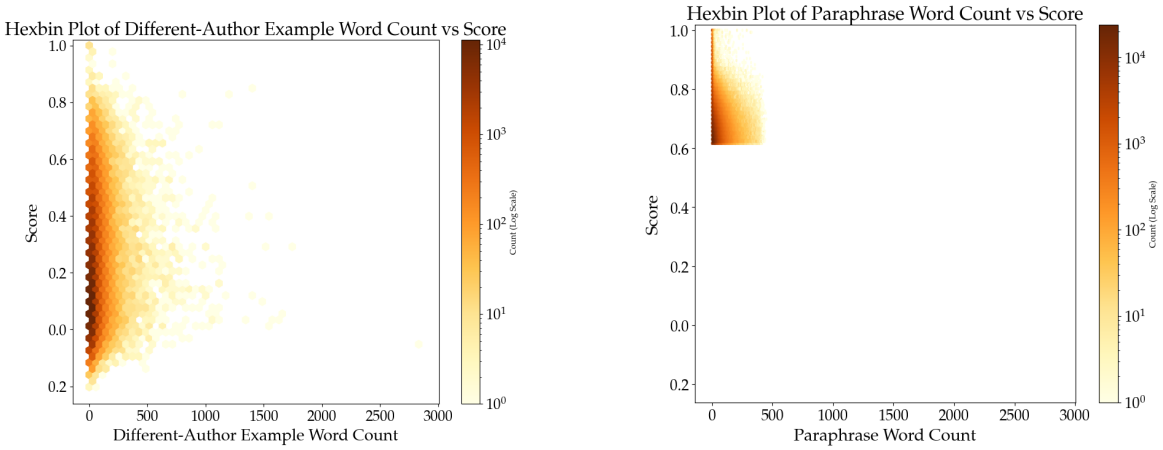
Figure 4.6: Side-by-side comparison between the frequency distribution for the first 10,000 different-author examples used by Wegmann, Schraagen and Nguyen (2022) (left) and by me (right), sorted by occurrences. The x-axis represents the index of the sorted different-author examples, while the y-axis represents the number of occurrences of each different-author example. For a fair comparison, both plots have the same x and y limits.

are too long, and (ii) the previously mentioned notion that it is easier to find semantically similar examples for shorter utterances. Second, despite the higher word counts, their approach still favors shorter sentences. However, this does not appear to be correlated with the similarity score (as is the case in my dataset) but rather seems to be a general property of language, where shorter sentences are more common.

Lastly, an additional important aspect to consider is the number of conversations from which the different-author examples are sampled. This is crucial because if most of my different-author examples were drawn from the same conversation, it would suggest that my approach for content control is actually quite similar to the approach taken by Wegmann, Schraagen and Nguyen (2022). In their approach, all different-author examples are from the same conversation as the anchor. However, in my approach, only 320,824 out of 2,995,445 examples, or 10.71%, are from the same conversation. This is a substantial difference and suggests that my approach is not merely a different conversation sampling method.

4.1.5 *Conclusion*

In this analysis, I've highlighted the important differences in the sampling approach used in my dataset compared to the method proposed by Wegmann, Schraagen and Nguyen (2022). While their approach tends to result in fewer similar different-author examples, my similarity-based approach provides a more focused context, which could be advantageous depending on the task of style representation.
The overrepresentation of shorter, semantically similar paraphrases might initially appear



(a) Hexbin plot comparing the word count of the different-author examples and their corresponding cosine similarity scores used by Wegmann, Schraagen and Nguyen (2022).

(b) Hexbin plot comparing the word count of the different-author examples and their corresponding cosine similarity scores used by me.

Figure 4.7: Side-by-side comparison between hexbin plots comparing the word count of the different-author examples and their corresponding cosine similarity scores. This comparison is between the data used by Wegmann, Schraagen and Nguyen (2022) (left) and me (right). The x-axis represents the word count in the different-author examples, and the y-axis represents the score associated with the different-author examples. The color of the hexagons represents the count of example-score pairs that fall into the area, with darker colors indicating higher counts, plotted on a log scale. This plot visualizes the density and distribution of the data, as well as any potential correlations between example word count and the score assigned by the model. For a fair comparison, both plots have the same x and y limits.

beneficial for the specific task at hand, but an overemphasis on these types of paraphrases can potentially lead to an unintended consequence: they might dominate the learning process to such an extent that the model fails to adequately capture the essence of complex and distinctive phrases. Similarly, the phenomenon of lower-scoring paraphrases being longer could induce a bias in the model to perform better on simpler paraphrases, potentially neglecting complex counterparts.

Moreover, the presence of both lower-scoring and frequently used paraphrases in the dataset, combined with a diverse range of writing styles for the same utterances, results in a unique structure. This could be advantageous for training, as the frequent paraphrases, which represent common structures and patterns in the language, provide the model with a wide array of base scenarios to learn from. Essentially, these common structures serve as the groundwork upon which the model can learn more intricate patterns. However, there is a potential drawback to this distribution pattern: it may result in the overrepresentation of certain scenarios, skewing the model's understanding and leading to an imbalance in representation.

The degree to which these characteristics impact the performance of the model, and the manner in which they do so, will require further exploration. As such, the potential upsides and downsides of these factors will only become fully evident upon the actual evaluation of the model results. By proceeding with this training and subsequent test-

ing, a clearer understanding of the effectiveness of this paraphrase-based method for the sampling of the different-author examples can be achieved.

4.2 OVERVIEW OF EXPERIMENTS

In this section, I provide an overview of the entire experimental process, summarizing the main steps undertaken and pointing toward the sections where each stage is described in more detail. The experiments, as visualized in Fig. 4.8, were carried out in a step-wise fashion, where each successive model builds on the results of the previous one.

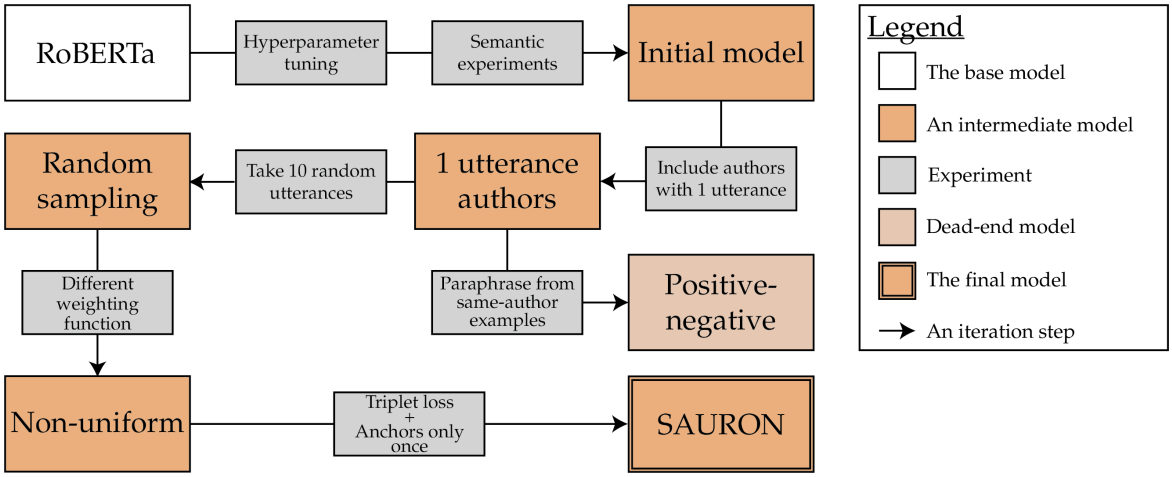


Figure 4.8: A flowchart summarizing the various different experiments and resulting models presented in the following sections of this thesis.

1. **Initial model:** The experiment started with the hyperparameter tuning phase, resulting in the creation of an initial model. This model employed the most suitable hyperparameters identified during the tuning phase. Moreover, it was tested with varying proportions of semantically similar different-author examples, with a 100% proportion proving to be optimal. For a comprehensive description of this initial model and the hyperparameter tuning process, please refer to Section § 4.3.

2. **1 utterance authors:** The initial model only included authors with at least two utterances, since it is also possible to sample same-author examples for those authors. In this next step, I adapted the data such that authors with only a single utterance could also be considered for the different-author examples. The motivation for this step was to increase the range of writing styles by including substantially more authors. More information about this model, and the modifications made, can be found in § 4.4.

3. **Positive-negative:** To experiment with the effectiveness of the training task setup, the third step involved a serious modification to the data sampling method for the different-author examples. Instead of finding the paraphrases of the anchor utterances, I instead mined them based on the same-author examples. An example of

the new training task is displayed in Fig. 4.10b, and detailed discussions about this model are provided in § 4.5.

4. **Random sampling:** To improve the results further after finding the results of the third step to be inferior to previous iterations, the training data was diversified using a random sampling method. This involved randomly selecting utterances from each author, rather than the first 10 chronological — as was the case for the previous experiments. The motive for this experiment was to increase the average number of topics for each author, which increase the diversity of writing styles even further. The random sampling model is thoroughly discussed in § 4.6.

5. **Non-uniform:** To improve the overall variety of paraphrases, a non-uniform sampling method to sample the different-author examples was applied to further change the model's performance. The intricacies of the non-uniform model and the motivation behind its creation are described in § 4.7.

6. **SAURON:** The final step involved the creation of the Stylistic AUthorship RepresentatiON (SAURON) model, which made two significant changes inspired by the approach of Wegmann, Schraagen and Nguyen (2022). The *MultipleNegativesRankingLoss* function was replaced with the Triplet loss function, aimed to foster a more nuanced and balanced representation by ensuring a clearer distinction between the anchor's relationships with positive and negative examples. Furthermore, the data sampling strategy was altered to deploy each anchor only once, necessitating a shift in the model's learning to glean stylistic information from a less populated, albeit more distinct, set of examples. The motivation for this experiment was to explore the possibility of achieving a more optimized balance between style and content control by integrating elements from the model developed by Wegmann, Schraagen and Nguyen (2022). An extensive discussion of the SAURON model, its development, and its performance can be found in § 4.8.

A comprehensive summary of the results from each of these stages is provided in Table 4.3. This iterative process has allowed for continuous refinements to the model, and this section serves to provide an overview of how each step contributed to the final results. In the following sections, the results of each of these models will be analyzed in more depth to provide a better understanding of how each modification impacts the model's performance.

4.3 INITIAL EXPERIMENTS

As a starting point for my experiments, I used a dataset structured in a triplet setup, with each input example consisting of an anchor, a positive (same-author), and a negative (different-author) text. The anchor text is written by author A, the positive example is another text that is also written by author A, and the negative example is a text written by another author B. To enhance the challenge of distinguishing between authors, the negative example is selected based on its semantic similarity to the anchor text, as measured by

	AV task		CAV task		Formal		Complex		Nb3r		C'tion	
Model	0%	100%	0%	100%	Original	o-c	Original	o-c	Original	o-c	Original	o-c
RoBERTa	0.58	0.50	0.63	0.57	**0.83**	0.09	**0.73**	0.01	**1.00**	0.00	0.94	**0.13**
Wegmann et al. (2022)	0.67	0.63	0.73	0.68	**0.83**	**0.70**	0.58	**0.27**	0.56	0.03	0.96	0.02
Initial model	0.72	0.60	0.80	0.64	0.81	0.48	0.59	0.04	0.65	**0.06**	0.95	0.00
Adjusted authors	0.75	0.62	0.82	0.68	0.78	0.47	0.57	0.04	0.72	0.03	**1.00**	0.00
Positive-negative	**0.78**	0.55	**0.85**	0.55	0.76	0.16	0.58	0.00	0.63	0.04	**1.00**	0.00
Random sampling	0.73	0.61	0.83	0.67	0.81	0.48	0.55	0.05	0.61	0.03	**1.00**	0.00
Non-uniform	0.75	0.62	0.83	0.68	0.80	0.48	0.58	0.04	0.67	0.04	**1.00**	0.00
SAURON	0.64	**0.64**	0.71	**0.73**	0.78	0.68	0.55	0.25	0.49	0.04	0.95	0.09

Table 4.3: Table comparing the performance of the RoBERTa base model, the model from Wegmann, Schraagen and Nguyen (2022), and all models presented in this thesis on the AV and CAV tasks using 0% and 100% semantically similar negative examples, as well as the STEL framework. The Formal, Complex, Number substitution and Contraction task columns are subdivided into Original (standard STEL dataset) and o-c (STEL-or-content dataset) subcolumns. The values with the highest accuracy in each column are reported in **bold**. The underlined values correspond to the highest accuracy in each column for the models that I present throughout this chapter.

cosine similarity scores. Whenever possible, I opted for the highest-scoring semantically similar text by a different author as the negative example. Furthermore, I sampled the texts chronologically to capture the natural evolution of an author's writing style over time.

This approach was used across the initial experiments. In Section § 4.3.2, I subsequently experiment with the proportion of different-author examples that are semantically similar. For more details on the collection method and the dataset description, refer to Section § 3.1.

4.3.1 *Hyperparameters*

To establish a baseline for the rest of my research, I experimented with different loss functions, learning rates, and the number of training epochs, fine-tuning these hyperparameters to optimize the model's performance in capturing stylistic nuances. A brief justification for the choices made regarding these hyperparameters is provided below, and detailed results of these experiments can be found in Appendix A.

4.3.1.1 *Loss function*

I first compared the performance of two different loss functions: *MultipleNegativesRankingLoss* (MNRL)[2] (Gao, Yao and Chen, 2021) and the *Contrastive* loss[3] (Hadsell, Chopra and LeCun, 2006). These loss functions were chosen because both are designed for contrastive learning tasks, which aim to learn representations that distinguish between similar and dissimilar data points.
The MNRL loss is originally designed for learning semantic sentence embeddings. It considers other inputs from the same batch, thereby enabling the model to learn from

2 https://www.sbert.net/docs/package_reference/losses.html#multiplenegativesrankingloss
3 https://www.sbert.net/docs/package_reference/losses.html#contrastiveloss

multiple positive and negative pairs simultaneously. Essentially, for a given positive pair, it treats all other negatives in the batch as valid negatives, which adds diversity to the negative samples.
The Contrastive loss, on the other hand, is a more general-purpose contrastive loss function that aims to push positive pairs close together and negative pairs apart in the embedding space. Comparing these two loss functions helps to identify the most suitable one for learning stylistic embeddings in the context of the Authorship Verification task.
Comparing the performance of these two loss functions provides insights into the most suitable loss function for learning stylistic embeddings. To ensure a fair comparison, the model was trained using each loss function with the number of epochs fixed at 5 and all other hyperparameters kept consistent.

The results reported in Table A.1 revealed that the model trained with *MultipleNegativesRankingLoss* outperformed the one trained with the *Contrastive* loss on almost all tasks. The differences in performance ranged from 2 to 9 percentage points on the AV and CAV tasks, and 0 to 6 percentage points on the STEL tasks. Therefore, MNRL was chosen as the preferred loss function for the subsequent experiments.

4.3.1.2 *Learning rate*

After determining the loss function, I experimented with different learning rates, including 1e-5, 2e-5, 3e-5, and 4e-5, using the *AdamW* optimizer. To ensure a fair comparison, I used the best settings from the previous experiments, which included the optimal number of epochs and the most effective loss function.
After analyzing the results from Table A.2, I found that the performance did not convincingly change when varying the learning rate. As such, I decided to keep the default learning rate for RoBERTa (2e-5) for the subsequent experiments.

4.3.1.3 *Number of epochs*

In the third part of the hyperparameter tuning, I experimented with various values of epochs (2, 3, 4, 5, and 6) to determine the optimal number. During these experiments, all other hyperparameters were kept constant: the learning rate was set to the default value for RoBERTa (2e-5), the batch size was 8 (the maximum that could fit into the GPU), and the loss function used was *MultipleNegativesRankingLoss*.

The results from Table A.3 showed that using 5 epochs provided the best performance, as the accuracy on the AV task continued to improve on the validation data for up to 5 epochs. However, when training for more than 5 epochs, the accuracy started to decrease, likely due to overfitting. Consequently, a maximum of 5 epochs were chosen as the optimal value for this hyperparameter. It is important to note, however, that the model with the best validation accuracy is used. This means that the model is not guaranteed to train for 5 epochs.

4.3.1.4 *Batch size*

After establishing the initial hyperparameters, I proceeded to conduct further experiments, focusing specifically on the batch size. The batch size plays an important role in the training process of machine learning models, as it determines the number of examples used to compute the gradient for each update step. Smaller batch sizes can lead to a noisier gradient, which can help the model escape local minima, while larger batch sizes provide a more accurate estimation of the gradient, leading to more stable and consistent training.

In the context of the *MultipleNegativesRankingLoss* function, the choice of batch size has a particularly meaningful impact on the model's ability to learn meaningful style representations. This is because the loss function operates by comparing the similarity scores of positive and negative examples within a batch. A larger batch size provides more negative examples for the model to learn from, potentially leading to better discrimination between different writing styles. However, larger batch sizes can also increase the computational requirements and memory usage, which may become a limiting factor in the training process.
In contrast, some other loss functions, such as the cross-entropy loss used in classification tasks, may be less sensitive to batch size changes, as they compute the loss based on a single correct class label for each example rather than relying on the relationships between multiple examples within a batch.

After experimenting with a few different values for the batch size (2, 4, 8), I found that a batch size of 8 results in the best performance (see Table A.4). This result is generally not surprising, as larger batch sizes often lead to better generalization performance in deep learning models due to more accurate estimates of the gradient during training. Furthermore, a batch size of 8 coincides with the maximum size that could fit into the GPU used in these experiments. Thus, this further reinforces the choice of 8 as the optimal batch size for the subsequent phases of this study.

4.3.2 *Impact of random different-author sampling*

In this subsection, I investigate the impact of randomly sampling different-author examples as opposed to using semantically similar samples. I compare the performance of the model trained with different proportions of semantically similar different-author examples (100%, 50%, and 0%) to understand the importance of using semantically similar samples for learning writing style representations in the context of the Authorship Verification and Contrastive Authorship Verification tasks.

To conduct this experiment, I used the same anchor sentences and positive examples as in the previous experiments. However, I varied the different-author examples depending on the proportion of semantically similar samples used. For the 100% semantically similar setup, I used the best model from the last step of Section § 4.3.1. For the 50% semantically

similar setup, half of the different-author examples were semantically similar to the anchor sentences, while the other half were sampled randomly. Finally, for the 0% semantically similar setup, I used randomly sampled different-author examples without considering their semantic similarity to the anchor sentences.

It's important to note that I did not perform hyperparameter tuning for each individual setup (i.e., 100%, 50%, and 0% semantically similar). Instead, the hyperparameters were kept constant across the different configurations. This decision was guided by practical considerations, as extensive hyperparameter tuning across different setups would substantially increase computational time and complexity. However, it should be noted that this approach may not yield the optimal performance for each individual setup, as different levels of semantic similarity might benefit from different hyperparameter configurations. This limitation is a trade-off that was accepted in the interest of broader and more efficient experimentation.

4.3.2.1 *Results*

Training task	**AV task**		**CAV task**		**Formal**		**Complex**		**Nb3r**		**C'tion**	
% Semantic	0%	100%	0%	100%	Original	o-c	Original	o-c	Original	o-c	Original	o-c
0%	0.59	0.53	**0.82**	0.53	0.76	0.24	0.56	0.01	**0.65**	0.05	**0.99**	**0.00**
50%	0.64	0.55	**0.82**	0.61	0.79	0.34	0.58	0.01	**0.65**	**0.06**	0.98	**0.00**
100%	**0.72**	**0.60**	0.80	**0.64**	**0.81**	**0.48**	**0.59**	**0.04**	**0.65**	**0.06**	0.95	**0.00**

Table 4.4: The table above presents the experiments' results on the impact of random different-author sampling on model performance, with accuracies displayed as percentages. Columns for the AV Task and CAV Task show results for 0% and 100% semantically similar conditions, respectively. The Formal, Complex, Number substitution and Contraction task columns are subdivided into Original (standard STEL dataset) and o-c (STEL-or-content dataset) subcolumns. The values with the highest accuracy in each column are reported in bold.

When examining the results of the AV and CAV tasks, I find that the model trained with 100% semantically similar different-author examples consistently achieves the highest performance in the more challenging settings — the 100% AV and CAV tasks, and all the STEL-or-content tasks. This reinforces my belief that training with semantically similar different-author examples effectively equips the model to differentiate writing styles under semantically demanding circumstances. However, the benefits of semantically similar different-author examples extend beyond the most challenging scenarios. When looking at the 50% semantically similar condition, it's noticeable that this model generally outperforms the 0% semantic model and closely follows the performance of the 100% semantic model. This indicates that a higher share of semantically similar different-author examples is positively correlated with better performance.
Interestingly, for the simpler CAV task, where different-author examples are expected to have a low semantic resemblance to the anchor, the model trained without any semantically similar different-author examples outperforms the other models. This result might hint at the particularity of the training scenario: the model seems to perform better

when its training conditions more closely align with the testing conditions. In this case, both training and testing used contrastive and non-semantically similar different-author examples. It highlights the potential advantage of matching the training environment with the anticipated testing conditions for optimal performance.

Looking at the performance on the STEL framework dimensions, the impact of the proportion of semantically similar different-author examples appears to be less pronounced. However, the 100% semantically similar model does exhibit marginally better performance in the STEL-or-content (o-c) tasks. The accuracy on the "Formal" dimension doubles from the 0% training task to the 100%, while the "Complex" and "Number Substitution" dimensions improve by 3 percentage points and 1 percentage point, respectively. This suggests that incorporating these semantically similar different-author examples during training can offer an edge in these tasks, which - in turn - might indicate improved stylistic representations.

4.3.3 *Summary of findings and baseline comparison*

Based on the previous experimentation, it was determined that a batch size of 8 works best for the fine-tuned RoBERTa model using the *MultipleNegativesRankingLoss* function that has a learning rate of 2e-5 using the *AdamW* optimizer. Additionally, upon reflection of the results in Table 4.4, I also opted to fully sample the negative examples from the paraphrases of the anchor sentence. With these hyperparameters established in Section § 4.3.1, I now present a comparison between the best model I have created so far and the chosen baseline model from Wegmann, Schraagen and Nguyen (2022). The descriptions and motivations of the chosen baseline models can be found in Section § 3.3. A table summarizing the results can be found in Table 4.5.

	AV task		**CAV task**		**Formal**		**Complex**		**Nb3r**		**C'tion**	
Model	0%	100%	0%	100%	Original	o-c	Original	o-c	Original	o-c	Original	o-c
RoBERTa	0.58	0.50	0.63	0.57	**0.83**	0.09	**0.73**	0.01	**1.00**	0.00	0.94	**0.13**
Wegmann et al. (2022)	0.67	**0.63**	0.73	**0.68**	**0.83**	0.70	0.58	**0.27**	0.56	0.03	**0.96**	0.02
Initial model	**0.72**	0.60	**0.80**	0.64	0.81	0.48	0.59	0.04	0.65	**0.06**	0.95	0.00

Table 4.5: Table comparing the performance of the RoBERTa base model, the model from Wegmann, Schraagen and Nguyen (2022), and the proposed model ("Initial model") on the AV and CAV tasks using 0% and 100% semantically similar different-author examples, as well as the STEL framework. The table shows the results for both the original tasks and the STEL-or-content (o-c) tasks. The values with the highest accuracy in each column are reported in bold.

Upon analysis of the results, a few things stand out. Firstly, it is evident that the proposed model consistently outperforms the RoBERTa base model in both AV and CAV tasks. This suggests that the fine-tuning process and contrastive learning approach have been effective in capturing writing styles for the Authorship Verification task. Interestingly, while the proposed model surpasses previous work in the relatively simpler AV and CAV tasks, which use randomly sampled different-author examples, it falls short when compared to

the model by Wegmann, Schraagen and Nguyen (2022) in the more challenging setting of the AV task with 100% semantically similar different-author examples. This is an interesting observation, considering that their model was not explicitly trained on this task. However, it is important to consider the possibility that their model may have been trained on some of the same data used in the testing set for the evaluation. In order to further investigate this impact, the proposed model was also evaluated on the AV and CAV test sets from their paper. Those results can be found in Table 4.6.

	AV task	**CAV task**
Wegmann et al. (2022)	**0.63**	**0.68**
Initial model	0.61	**0.68**

Table 4.6: Table comparing the performance of the model from Wegmann, Schraagen and Nguyen (2022) and my initial model on the AV and CAV tasks, using the test sets from their paper. This testing task features their conversation-level content control. The values with the highest accuracy in each column are reported in bold.

A quick analysis of the results presented in Table 4.6 reveals that the proposed model performs competitively when evaluated on the AV and CAV tasks from Wegmann, Schraagen and Nguyen (2022). In the AV task, the model from Wegmann, Schraagen and Nguyen (2022) demonstrates slightly superior performance. However, in the CAV task, both models achieve equal performance. It's worth noting that these tasks apply conversation-level content control, a feature that was a part of the training strategy for the model from Wegmann, Schraagen and Nguyen (2022). Even so, the proposed model, which was not explicitly trained in this way, manages to perform on par in the CAV task and very close in the AV task.

The proposed model's performance on the STEL dimensions is mixed. While it outperforms the RoBERTa base model in some cases, it does not consistently surpass the results of the Wegmann, Schraagen and Nguyen (2022) model, suggesting that there is still room for improvement in capturing stylistic features across various dimensions. Additionally, the performance of the proposed model on STEL-or-content (o-c) tasks appears to be weaker than on the original tasks, indicating that the model is still somewhat reliant on content features rather than purely learning stylistic features.

4.4 ADJUSTING AUTHOR SAMPLING

From the aforementioned insights, I identified potential areas of improvement, such as increasing the amount of data, the number of conversations the data was sampled from, or both. Upon closer examination of the dataset, I discovered that almost all conversations were already included (59,962 used out of 60,000 total). Therefore, expanding the number of conversations was not a viable option without creating an entirely new corpus.

However, in the previous experiments, I had intentionally left out authors with only

one utterance, as it was not possible to create anchor-positive pairs for them. By revisiting this decision, I found a potential avenue to enhance the dataset: using the single-utterance authors as paraphrases and negative examples. This approach would not only increase the amount of data but also introduce more authors and, possibly, a wider variety of writing styles into the dataset.

To prevent the overrepresentation of some authors and the underrepresentation of others, I also limited the maximum number of utterances from each author. In this experiment, I took the first 10 (chronologically ordered) utterances from each author, aiming to create a more balanced representation of writing styles.
The result of this different sampling strategy is almost similar to that of Table 3.2, but changes after the "1 utterance" pre-processing step. Instead of removing 474,339 utterances and authors, I removed 1,431,857 utterances, without removing any authors. After removing the utterances that were too long, 3,080,638 utterances by 1,011,947 authors were left over. Although this decreased the number of utterances by 23,7% (from 4,038,747 to 3,080,638), the number of authors increased from 538,199 to 1,011,947 — an increase of 88%. By limiting the number of utterances per author to 10 while increasing the number of authors, I prevented the problem of overrepresentation of some authors, while possibly increasing the diversity of writing styles.

4.4.1 *Results*

In this experiment, I adapted the initial model setup, where optimal hyperparameters were determined and 100% semantically similar different-author examples were found to be most effective. The main modification involved adjusting the data to include authors with only a single utterance and limiting the dataset to the first 10 utterances from each author. This model is referred to as the "1 utterance authors" model.

The table below compares the performance of the RoBERTa base model, the model from Wegmann, Schraagen and Nguyen (2022), the initial model, and the new "1 utterance authors" model on the AV and CAV tasks, as well as the STEL framework dimensions formal/informal, complex/simple, number substitution, and contraction tasks.

	AV task		**CAV task**		**Formal**		**Complex**		**Nb3r**		**C'tion**	
Model	0%	100%	0%	100%	Original	o-c	Original	o-c	Original	o-c	Original	o-c
RoBERTa	0.58	0.50	0.63	0.57	**0.83**	0.09	**0.73**	0.01	**1.00**	0.00	0.94	**0.13**
Wegmann et al. (2022)	0.67	**0.63**	0.73	**0.68**	**0.83**	**0.70**	0.58	**0.27**	0.56	0.03	0.96	0.02
Initial model	0.72	0.60	0.80	0.64	0.81	0.48	0.59	0.04	0.65	0.06	0.95	0.00
1 utterance authors	**0.75**	0.62	**0.82**	**0.68**	0.78	0.47	0.57	0.04	0.72	0.03	**1.00**	0.00

Table 4.7: Table comparing the performance of the RoBERTa base model, the model from Wegmann, Schraagen and Nguyen (2022), the model from § 4.3.3 ("Initial"), and the new proposed model ("1 utterance authors") on the AV and CAV tasks using 0% and 100% semantically similar different-author examples, as well as the STEL framework. The table shows the results for both the original tasks and the STEL-or-content (o-c) tasks. The values with the highest accuracy in each column are reported in bold.

The results in Table 4.7 show that the "1 utterance authors" model demonstrates improved performance in the AV and CAV tasks compared to the old proposed model, especially in the 0% semantically similar different-author examples condition. When comparing the old and new proposed models, it is also noteworthy that the performance on the STEL framework dimensions remains relatively consistent. This observation suggests that the alterations made to the dataset and experimental setup did not negatively impact the model's ability to capture stylistic features in the text, but also had no positive impact. It is possible, however, that the performance on the STEL framework is relatively constant between the old and new proposed models for several other reasons. Firstly, it could be that the STEL framework may not be fully representative of the stylistic nuances present in the data, causing the models to not exhibit notable differences in performance. Secondly, the limited number of examples for each dimension ($n = 815$ for the formal and complex dimensions, and $n = 100$ for number substitution and contraction) in the dataset could lead to a lack of variability in the model's learning of these dimensions. A third reason could be that the dimension examples in the STEL framework may not translate well to the full diversity of writing styles, as they might not capture the full range of stylistic differences between authors. The following subsection will delve into a manual inspection of several examples from the STEL-or-content tasks to further our understanding of the specific challenges faced by the models, as well as to identify potential areas of improvement in the way these tasks are designed. This closer examination could offer valuable insights into the intricate complexities of style representation and disentanglement.

4.4.2 *STEL-or-content analysis*

The fact that the STEL framework's dimension examples may not represent the full spectrum of stylistic distinctions across authors is best illustrated in a figure. Fig. 4.9 shows an example from the formal dimension on the STEL-or-content task that the new proposed model got wrong. It can be argued that, although the style also differs from formal to informal, there are more stylistic differences between the Anchor and Utterance 2, than between the Anchor and Utterance 1. It is thus not surprising that the cosine similarity that the model has calculated between A and U1 (0.707) is much higher than between A and U2 (0.363).

Upon manual inspection of the mistakes that the model made on the STEL-or-content task, it becomes clear that this is a problem for more of the style dimensions. This is particularly apparent in the number substitution task, where the majority of misclassifications are associated with serious stylistic alterations like the full capitalization of utterances, varied punctuation use, or considerable sentence length disparities. The text pair initially labeled as "same style" with the highest cosine similarity score obtained a value of 0.51. Conversely, the text pair initially deemed as "not same style" achieved an exceptional cosine similarity score of 0.99. This discrepancy highlights potential challenges with the dataset. For instance, sourcing data from Reddit might not be the most conducive for the

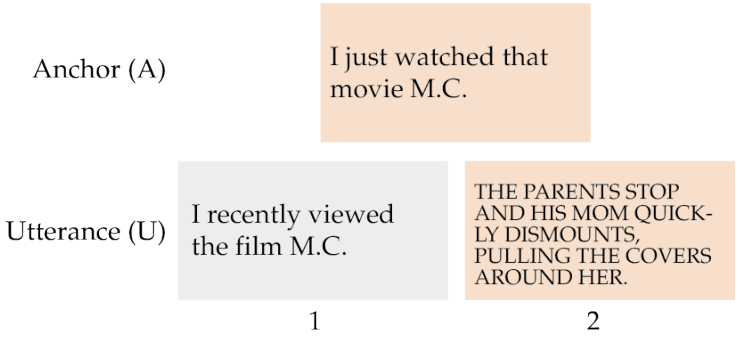


Figure 4.9: One of the STEL-or-content task instances on the formal/informal dimension. In this case, the ground truth is that the Anchor and Utterance 2 are written in the same style. My model assigns a cosine similarity score of 0.363 to the Anchor and U2, and a score of 0.707 to the Anchor and U1.

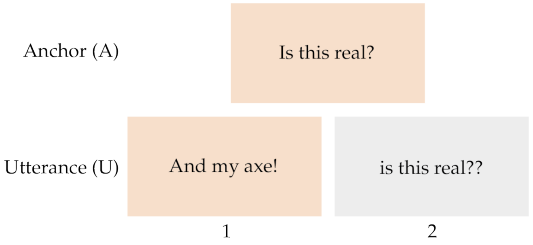
number substitution task, as the platform generally exhibits a relatively "clean" language style. Consequently, this type of writing style is underrepresented in the training data. Furthermore, the examples from this dimension often seem to involve other style changes, indicating a probable overlap of style dimensions.
The contraction dimension presents similar issues. The stylistic differences between utterances labeled as "different style" are often marginal when compared to the corresponding "correct" sentence, which brings the appropriateness of categorizing this as a separate dimension into question. This observation suggests that the style dimensions might not be mutually exclusive, and a more nuanced approach may be necessary to distinguish stylistic changes more effectively.
This problem is less visible in the other dimensions but it is still prevalent. For instance, approximately 25% of the errors made on the formal and complex dimensions can be attributed to factors like the ones exhibited in Fig. 4.9 and differences in casing. A substantial proportion of the remaining errors appears to stem from the model's inability to detect the sometimes subtle variations in writing style. In these instances, the model appears to prioritize content over style, possibly due to the influence of the underlying RoBERTa architecture. Trained with a strong focus on semantic understanding, RoBERTa might inadvertently sway the fine-tuned model towards content-based decisions, even when style is the defining factor. Even minute shifts in style can significantly alter the text's tone, register, or complexity. However, these nuanced changes often seem to evade the model's comprehension, pointing to a potential area of improvement.

Interestingly, the Wegmann, Schraagen and Nguyen (2022) model performs well on the formal and complex dimensions of the STEL framework, which may be due to the model's training process that includes controlling for content by sampling different-author examples from the same conversation. By choosing different-author examples within the same conversation, the model is exposed to texts with related content, and as such, it is prompted to differentiate authors based on variations in writing style rather than

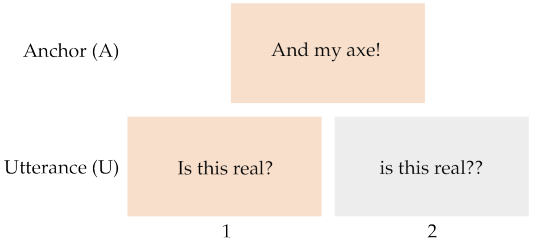
content. This approach may allow their model to better learn stylistic features that are more aligned with the STEL dimensions.

4.5 EXPERIMENTATION WITH NEGATIVE EXAMPLES

In light of the previous discussions, another potential approach to consider is selecting the different-author example as a paraphrase of the same-author example, rather than the current method of choosing the different-author example as a paraphrase of the anchor. Examples of the old and new approaches are given in Fig. 4.10a and Fig. 4.10b, respectively.
Intuitively, this alternative approach would make the contrastive task more challenging for the model, as the content of both the positive and negative examples would be the same. Consequently, the model would be forced to focus more on capturing stylistic differences between authors, as relying on content differences would no longer be a viable strategy for distinguishing between positive and negative examples. This modification could potentially lead to a more robust model for writing style representation and better performance on tasks related to Authorship Verification and style modeling.



(a) An input example where the different-author utterance is sampled as a paraphrase of the **anchor** sentence. This is the "old" approach.



(b) An input example where the different-author utterance is sampled as a paraphrase of the **same-author** utterance. This is the approach experimented with in this section.

Figure 4.10: Comparison of input examples for the Contrastive Authorship Verification task with the old (left) and new (right) sampling approach. In both scenarios, Utterance 1 and Utterance 2 correspond to the same-author and different-author examples, respectively.

4.5.1 *Results*

In this setup, I tweaked the "1 utterance authors" model as described in Section § 4.4. The alteration involved the selection of the different-author example which, in this case, is a paraphrase of the same-author example instead of the anchor. This new model variant is termed the "Positive-negative" model.

Upon examining the results presented in Table 4.8, I find them a bit surprising. When compared to the previous experiments, the model underperforms on almost all tasks, even falling short of the performance of the base RoBERTa model on the 100% semantically similar CAV task. This discrepancy is quite intriguing, especially considering

	AV task		CAV task		Formal		Complex		Nb3r		C'tion	
Model	0%	100%	0%	100%	Original	o-c	Original	o-c	Original	o-c	Original	o-c
RoBERTa	0.58	0.50	0.63	0.57	**0.83**	0.09	**0.73**	0.01	**1.00**	0.00	0.94	**0.13**
Wegmann et al. (2022)	0.67	**0.63**	0.73	**0.68**	**0.83**	**0.70**	0.58	**0.27**	0.56	0.03	0.96	0.02
1 utterance authors	0.75	0.62	0.82	**0.68**	0.78	0.47	0.57	0.04	0.72	0.03	**1.00**	0.00
Positive-negative	**0.78**	0.55	**0.85**	0.55	0.76	0.16	0.58	0.00	0.63	0.04	1.00	0.00

Table 4.8: Table comparing the performance of the RoBERTa base model, the model from Wegmann, Schraagen and Nguyen (2022), the previously proposed model ("1 utterance authors"), and the new proposed model ("Positive-negative") on the AV and CAV tasks using 0% and 100% semantically similar negative examples, as well as the STEL framework. The table shows the results for both the original tasks and the STEL-or-content (o-c) tasks. The values with the highest accuracy in each column are reported in bold.

the intuition behind the strategy: by choosing the different-author example as a paraphrase of the same-author example, the model should be pushed towards focusing more on stylistic elements, as the content should be virtually identical. However, this strategy seems to excel in the simpler setting of the AV and CAV tasks when there is no semantic similarity between the anchor and the different-author example. The model delivers the best performance overall for both the 0% semantically similar AV task and CAV tasks, which suggests that when the content of the different-author example differs greatly from the anchor, the model successfully leverages stylistic features to distinguish authors.
Taking into account these findings alongside the results from the STEL-or-content task, it seems plausible that the model potentially assigns greater importance to features that are not strictly stylistic for Authorship Verification. It is possible that the content or thematic elements of the text may play a substantial role in the model's decision-making process, influencing its ability to accurately verify authorship.

As for why this strategy may not have lived up to expectations on the 100% CAV and AV tasks, and the STEL tasks, it could be that the training task was simply too challenging. By forcing the model to differentiate authors based solely on stylistic elements, without any clear content-related cues to rely upon, the model might have struggled to learn meaningful stylistic representations. This increased difficulty may be reflected in the poor performance on tasks that require a nuanced understanding of stylistic differences between authors.
Another possibility worth considering is the quality and nature of the paraphrases used in the training data. As mentioned in § 4.1.5, one of the drawbacks of my paraphrase mining approach could be that the occurrence of shorter paraphrases is more prevalent than the use of longer ones. In other words: the model might simply not perform well on longer sentences, since it has seen less of those in the training data. Upon quick inspection of the sentences from the STEL dataset, it seems that those sentences tend to be longer than most of my best-scoring different-author examples. Given that in the STEL-or-content task, we also want to distinguish content from style using a paraphrase (with near-identical semantics), my model might run into even more problems. The combination of longer, semantically equivalent texts is not common in my data — instead, semantically equivalent texts tend to be shorter. The different-author example sampling

approach employed by Wegmann, Schraagen and Nguyen (2022) is relatively independent of utterance length, which might give them an edge. Their method allows them to sample longer, albeit sometimes less semantically similar sentences. Another benefit of their approach is that they can find a negative example for each of their positives — something I cannot do since I can not always find a semantically similar counterpart for an utterance in my data.

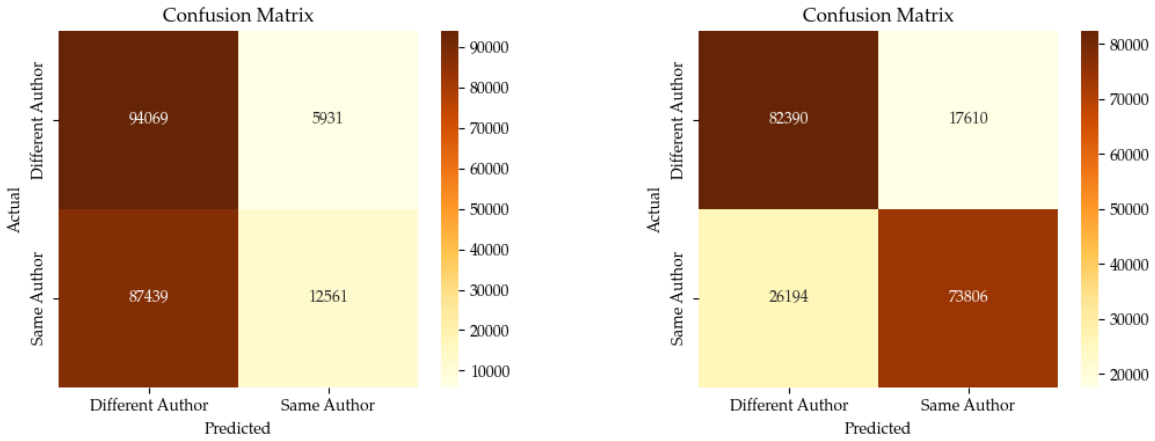
A potential way to verify some of these hypotheses would be to conduct an analysis of the model's errors on the AV task. By examining instances where the model failed to correctly classify authors, I could gain insights into the specific challenges that this training strategy posed. These investigations could potentially shed light on the limitations of the current strategy and provide direction for future experiments.

4.5.2 *AV error analysis*

Upon analyzing the confusion matrix from my model on the 100% Authorship Verification task (see Fig. 4.11a), a few key observations can be made. While the accuracy of the model is 0.55, slightly better than a random guess, the precision, recall, and F1-score are 0.60, 0.53, and 0.44, respectively. These metrics suggest significant room for improvement.

A closer look at the model's behavior, as illustrated in Fig. 4.11a, reveals it has a tendency to classify 90.8% of all examples as "different authors." Given that the class distribution is perfectly balanced (50-50), this differs quite a lot from the expectation of a more even prediction distribution. One might initially think the classification threshold for "same author" could be set too high. However, adjustments to this threshold based on the test set rather than the validation set did not considerably change the results, indicating the threshold might not be the main issue.
As mentioned in the previous subsection, a plausible explanation for this behavior could be that in the way this particular task is set up, the training process does not adequately help the model disentangle style from content. Consequently, the model may be overly reliant on content cues, which are not as prominent in the 100% task. As a result, adjusting the threshold might not yield substantial improvements because it does not address the underlying issue of the model's tendency to focus on content over style.

This conjecture is somewhat supported when examining the confusion matrix for the model on the 0% AV task, as depicted in 4.11b. The accuracy in this scenario notably improves compared to the more challenging task, with the "Positive-negative" model achieving the highest score on the 0% AV and CAV tasks out of all tested models so far (see Table 4.3). This success can likely be attributed to the training tasks enabling the model to focus effectively on content cues.
However, the model does not favor predicting "Different author" as frequently in this task, indicating its dependency on content for decision-making. When the task adjusts to make



(a) Confusion matrix of the results on the 100% AV task. The accuracy is 0.55, precision is 0.60, recall is 0.53, and F1-score is 0.44.

(b) Confusion matrix of the results on the 0% AV task. The accuracy is 0.78, precision is 0.78, recall is 0.78, and F1-score is 0.78.

Figure 4.11: Comparison of confusion matrices of the results on the Authorship Verification task with the 100% (left) and 0% (right) semantically similar different-author examples test sets.

content cues less reliable (as in the 100% AV task), the model's performance decreases considerably. This reinforces the idea that my training methodology may need further refinement to improve the disentanglement of style from content.
This dependency on content also sheds light on the low performance on the STEL-or-content tasks. As these tasks are designed to evaluate the model's ability to disentangle style from content, the model struggles to highlight its challenges in accurately identifying stylistic nuances independent of content.

4.6 RANDOMLY SELECTING UTTERANCES

Based on the conclusion that the "Positive-negative" model performs poorly at detaching style from content, I will revert back to the approach employed for my previous model ("1 utterance authors" — § 4.4). Another aspect I am considering based on the deliberations from the previous evaluations, is how the selection of utterances from each author could affect the model's performance. Until now, I have been using the first 10 utterances from each author, which are ordered chronologically. However, this approach could potentially limit the diversity of the training data in terms of the number of conversations included. The chronological ordering of utterances might result in a concentration of samples from a few conversations that occurred early in an author's contribution to the dataset. This limitation could potentially hinder the model's ability to capture the full breadth of an author's stylistic range, as the author's style might evolve across different conversations.

To address this concern, I adopt a strategy of selecting 10 random utterances from each author instead of the first 10. This approach would likely result in a more diverse set of utterances for each author and encompass a wider range of conversations. By increasing

the diversity of the training data in this way, I could potentially enable the model to better learn and generalize the stylistic characteristics of each author, which might improve its performance on the Authorship Verification and Contrastive Authorship Verification tasks, and subsequently also learn better style representations. I will explore this strategy in the following analysis.

The key change of this experiment is thus the adaption of the "1 utterance authors" model to select 10 random utterances from each author, rather than the first 10. This variant is referred to as the "Random sampling" model.

	AV task		**CAV task**		**Formal**		**Complex**		**Nb3r**		**C'tion**	
Model	0%	100%	0%	100%	Original	o-c	Original	o-c	Original	o-c	Original	o-c
RoBERTa	0.58	0.50	0.63	0.57	**0.83**	0.09	**0.73**	0.01	**1.00**	0.00	0.94	**0.13**
Wegmann et al. (2022)	0.67	**0.63**	0.73	**0.68**	**0.83**	**0.70**	0.58	**0.27**	0.56	**0.03**	0.96	0.02
1 utterance authors	**0.75**	0.62	0.82	**0.68**	0.78	0.47	0.57	0.04	0.72	**0.03**	**1.00**	0.00
Random sampling	0.73	0.61	**0.83**	0.67	0.81	0.48	0.55	0.05	0.61	**0.03**	**1.00**	0.00

Table 4.9: Table comparing the performance of the RoBERTa base model, the model from Wegmann, Schraagen and Nguyen (2022), the previously proposed model ("1 utterance authors"), and the new proposed model ("Random sampling") on the AV and CAV tasks using 0% and 100% semantically similar different-author examples, as well as the STEL framework. The table shows the results for both the original tasks and the STEL-or-content (o-c) tasks. The values with the highest accuracy in each column are reported in bold.

Table 4.9 compares the performance of various models, including the RoBERTa base model, the model proposed by Wegmann, Schraagen and Nguyen (2022), a previously proposed model referred to as "1 utterance authors," (§ 4.4) and my newly proposed model, "Random sampling." The results show a similar performance of the "Random sampling" model compared to the "1 utterance authors" model across all tasks. For the AV and CAV tasks, the model presents mixed results. More specifically, while the model is outperformed by the "1 utterance authors" and Wegmann, Schraagen and Nguyen (2022) models on the 100% tasks, it performs exceptionally well on the 0% tasks.
The "Random sampling" model shows similar performance on the formal original and o-c tasks compared to the "1 utterance authors" model, demonstrating the model's unchanged capability to capture these stylistic features. An interesting observation can be made for the number substitution task, where all the models presented still have lower performance than the RoBERTa model. This could potentially be due to the inherent complexity of the task, indicating areas for future model improvements. A further explanation of the underperformance on the STEL tasks compared to Wegmann, Schraagen and Nguyen (2022) has also been given in Section § 4.4.2.

Taken together, these results suggest that the "Random sampling" model offers a similar performance across different tasks and styles, which brings into question the utility of randomly sampling the utterances of each author. The limited impact of this strategy on the model's performance might potentially be explained by the fact that a serious portion of authors in the dataset has fewer than ten utterances. Consequently,

shuffling the utterances would only affect a relatively small subset of authors, leading to a minor overall impact on the performance across various tasks. However, it's important to note that for larger datasets, exploring and potentially employing this strategy is recommended. Random sampling of the author's utterances could have a bigger impact when applied to larger and more diverse sets of authors and utterances.

4.7 NON-UNIFORM SAMPLING OF PARAPHRASES

Building upon the learnings from the "Random sampling" model, the next area to explore involves addressing the issue of paraphrase overrepresentation. One of the concerns raised in Section § 4.1 was the possibility of an overrepresentation of the same paraphrases when always selecting the highest-rated paraphrase as the different-author example. To diversify the sampling and ensure a more representative coverage of various paraphrases, I propose a different sampling strategy in this subsection.
In this approach, I begin by selecting the top 10 highest-rated paraphrases for a given sentence. Instead of invariably picking the top-rated paraphrase, I introduce a weighted selection mechanism. The weights are inversely proportional to the square root of the rank of the paraphrase. Formally, the weight for a paraphrase with rank i is given in Equation 4.1.

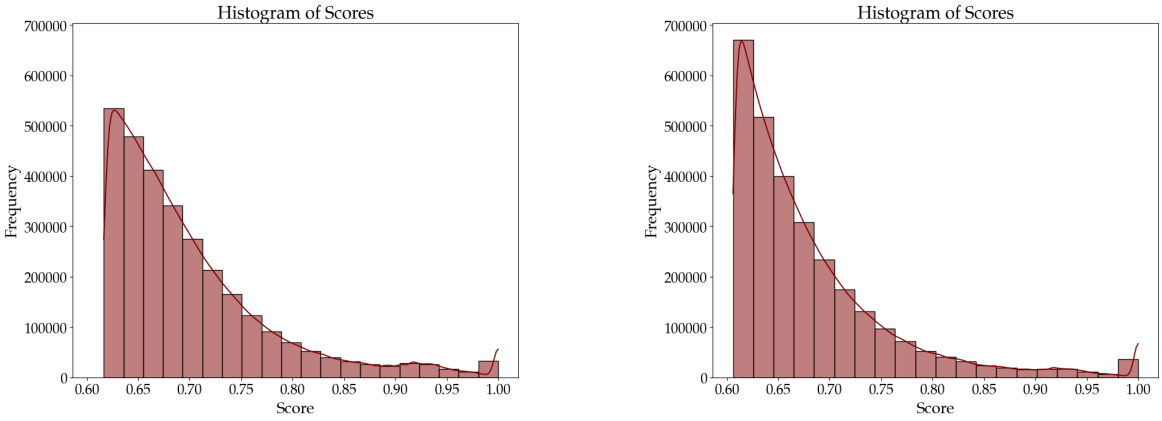
$$w_i = \frac{1}{1 + \sqrt{i}}$$

Equation 4.1: The weighting function for sampling paraphrases. The weights are subsequently normalized such that they sum up to 1. This makes the weight function a valid probability distribution over the top 10 paraphrases. Here, i is the rank of the paraphrase in the list of possible paraphrases and w_i is the probability of choosing paraphrase i.

The square root term $\sqrt{i}$, where i represents the rank of the paraphrase, ensures that the decay in the weights is not too steep. Consequently, the top-rated paraphrases are preferred, but not overwhelmingly so, thereby ensuring a decent representation from the entire top 10 list. By using this sampling strategy, the highest-ranked paraphrase is still more likely to be selected, but other high-ranking paraphrases are also given a reasonable chance. This non-uniform sampling method helps to diversify the different-author examples used during training and might improve the model's generalization by exposing it to a broader range of paraphrase variations.

Upon reviewing the score distribution in Fig. 4.12b, it is apparent that the distribution of paraphrase similarity scores has become even more skewed towards lower values, compared to the previous sampling method (Fig. 4.12a). This trend might initially appear concerning, but considering the distribution of scores from Wegmann, Schraagen and Nguyen (2022) depicted in Fig. 4.5a, it provides another perspective. Their model, despite using examples with substantially lower semantic similarity scores than mine,

demonstrates desirable performance. In other words, high semantic similarity scores are not necessarily indicative of superior performance. In fact, it presents an opportunity for an in-depth exploration of the relationship between model performance and semantic similarity. Through this experimental setup, I aim to uncover insights about how the semantic proximity of paraphrases may impact the effectiveness of style-learning models.



(a) Histogram of paraphrase scores showing the distribution of scores across the dataset. The scores range from a minimum of 0.62 to a maximum of 1.00, with a median of 0.68. The 25th and 75th percentiles are 0.64 and 0.73, respectively.

(b) Histogram of paraphrase scores showing the distribution of scores across the dataset. The scores range from a minimum of 0.61 to a maximum of 1.00, with a median of 0.66. The 25th and 75th percentiles are 0.63 and 0.71, respectively.

Figure 4.12: The cosine similarity score distribution before (left) and after (right) applying the new weighted sampling method. The kernel density estimate (KDE) is also plotted to give a smooth estimate of the score distribution. For a fair comparison, both plots have the same x and y limits.

In this variation, I adapted the "Random sampling" model outlined in Section § 4.6 by incorporating a weighted sampling function for paraphrase selection. I refer to this model as the "Non-uniform" model.

	AV task		**CAV task**		**Formal**		**Complex**		**Nb3r**		**C'tion**	
Model	0%	100%	0%	100%	Original	o-c	Original	o-c	Original	o-c	Original	o-c
RoBERTa	0.58	0.50	0.63	0.57	**0.83**	0.09	**0.73**	0.01	**1.00**	0.00	0.94	**0.13**
Wegmann et al. (2022)	0.67	**0.63**	0.73	**0.68**	**0.83**	**0.70**	0.58	**0.27**	0.56	0.03	0.96	0.02
Random sampling	0.73	0.61	**0.83**	0.67	0.81	0.48	0.55	0.05	0.61	0.03	**1.00**	0.00
Non-uniform	**0.75**	0.62	**0.83**	**0.68**	0.80	0.48	0.58	0.04	0.67	**0.04**	**1.00**	0.00

Table 4.10: Table comparing the performance of the RoBERTa base model, the model from Wegmann, Schraagen and Nguyen (2022), the previously proposed model ("Random sampling"), and the new proposed model ("Non-uniform") on the AV and CAV tasks using 0% and 100% semantically similar different-author examples, as well as the STEL framework. The table shows the results for both the original tasks and the STEL-or-content (o-c) tasks. The values with the highest accuracy in each column are reported in bold.

An analysis of the results, as shown in Table 4.10, reveals several noteworthy patterns. The newly proposed "Non-uniform" model shows the best performance in the "easy" AV and CAV tasks, with an accuracy of 0.75 and 0.83 on the 0% tasks respectively. However, this

model's performance in the more challenging tasks is either similar or inferior to that of the Wegmann, Schraagen and Nguyen (2022) model, particularly in the 100% AV and CAV tasks. Interestingly, while the performance of the "Non-uniform" model in the STEL tasks shows minor improvements over the previous "Random Sampling" model, particularly in the number substitution and contraction tasks, it does not exhibit the same trend in the STEL-or-content tasks. This suggests that while the "Non-uniform" model might be a bit better at generalizing style cues, it still struggles with disentangling style from content, a critical aspect of style-based tasks.

	Formal		**Complex**		**Nb3r**		**C'tion**	
Model	Original	o-c	Original	o-c	Original	o-c	Original	o-c
RoBERTa	0.83	0.09	**0.73**	0.01	**1.00**	0.00	0.94	**0.13**
Wegmann conversation	0.83	**0.70**	0.58	**0.27**	0.56	0.03	0.96	0.02
Wegmann topic	0.82	0.61	0.57	0.12	0.64	0.03	0.99	0.01
Wegmann none	**0.85**	0.50	0.56	0.04	0.59	**0.06**	0.98	0.00
Non-uniform	0.80	0.48	0.58	0.04	0.67	0.04	**1.00**	0.00

Table 4.11: Table comparing the performance of the RoBERTa base model, the three contrastive models from Wegmann, Schraagen and Nguyen (2022), and the proposed model ("Non-uniform") on the STEL framework. The table shows the results for both the original tasks and the STEL-or-content (o-c) tasks. The values with the highest accuracy in each column are reported in bold.

Upon analyzing table Table 4.11 - which compares the performance of the RoBERTa base model, the three contrastive models from Wegmann, Schraagen and Nguyen (2022), and the proposed model ("Non-uniform") on the STEL framework - it is intriguing to note that "Wegmann none", which does not employ any content control, surpasses the performance of the "Non-uniform" model on the STEL-or-content "formal" dimension. This seems counterintuitive, given the expectation that this model would be more focused on content than style, as discussed in Sections § 4.3 and § 4.5.
The comparatively less impressive result might be attributed to a variety of factors, including the different loss function used, the experimental setup which involves using each anchor only once, or other unidentified variables, such as potential biases in the dataset, or the balance between style and content in the training data.

On a related note, despite the slight improvements observed in the original STEL tasks, it's worth mentioning that neither of my models tested here perform exceptionally well on the STEL-or-content (o-c) tasks. This indicates that the challenge of distinguishing style from content continues to be a difficult problem for all tested models. This could be due to a number of reasons that I elucidated in § 4.4.1.

Overall, these results highlight the utility of my "Non-uniform" sampling approach in improving performance on certain tasks. However, they also underline the ongoing

challenges of training models to disentangle and accurately classify text based on stylistic nuances.

4.8 MOVING CLOSER TO WEGMANN, SCHRAAGEN AND NGUYEN (2022)

Reflecting on these results, it becomes evident that there's potential for further investigation. In the following section, I aim to harmonize my approach with that of Wegmann, Schraagen and Nguyen (2022). My models, which apply semantically-driven content control, exhibit similar performance on the STEL tasks as the model developed by Wegmann, Schraagen and Nguyen (2022) that does not implement any form of content control. This parallelism, although unexpected, introduces a compelling avenue for further investigation: what if I could harmonize my experimental approach with theirs, retaining my semantically similar approach, and simultaneously achieve a more optimized balance between style and content control?

To explore this hypothesis, I execute two key modifications inspired by their approach. First, I replace my original loss function, the *MultipleNegativesRankingLoss*, with the Triplet loss function employed by Wegmann, Schraagen and Nguyen (2022). The Triplet loss function uses an anchor, a positive example, and a negative example with the aim to maximize the representation space distance between the anchor and the negative example and simultaneously minimize the distance between the anchor and the positive example. Notably, the margin and cosine distance metric used are identical to those in the Wegmann, Schraagen and Nguyen (2022) model, set at 0.5. A description of the Triplet loss function is presented in Equation 4.2.

$$\mathcal{L}(a, p, n) = \max(0, \cos(a, p) - \cos(a, n) + \alpha)$$

Equation 4.2: The Triplet loss function, using the cosine similarity distance measure, where:
- a is the embedding of the anchor example
- p is the embedding of the positive example
- n is the embedding of the negative example
- cos is the cosine similarity distance function
- α is the margin between the positive and negative pairs

In the previous stages of my research, I employed the *MultipleNegativesRankingLoss* function, which primarily focuses on minimizing the distance between the anchor and positive examples. This approach guided the model to disentangle style from content by emphasizing the stylistic similarity between the anchor and the positive examples. While this method does consider negative examples, it does so in a more implicit manner by utilizing them to establish a boundary or a decision rule, rather than explicitly pushing the anchor away from these negative examples. Contrastingly, the Triplet Loss function adopts a more balanced strategy. It strives not only to bring the anchor and positive examples closer but also to widen the distance between the anchor and negative examples. This approach, consequently, fosters a more nuanced and balanced representation by ensuring a

clearer distinction between the anchor's relationships with positive and negative examples.

The second modification involves the data sampling strategy. Contrary to my original approach of generating as many (anchor, positive) pairs per anchor as feasible, their strategy deploys each anchor only once. This approach necessitates a shift in the model's learning, requiring it to glean stylistic information from a less populated, albeit more distinct, set of examples.

In this final variation of my thesis experiments, I transformed the "Non-uniform" model described in Section § 4.7 by introducing the Triplet loss function and employing each anchor only once. This version, referred to as the Stylistic AUthorship RepresentatiON (SAURON) model, encapsulates the culminating experiment that attempts to bridge the methodological gap between my original model and the approach outlined by Wegmann, Schraagen and Nguyen (2022).

	AV task		**CAV task**		**Formal**		**Complex**		**Nb3r**		**C'tion**	
Model	0%	100%	0%	100%	Original	o-c	Original	o-c	Original	o-c	Original	o-c
RoBERTa	0.58	0.50	0.63	0.57	**0.83**	0.09	**0.73**	0.01	**1.00**	0.00	0.94	**0.13**
Wegmann et al. (2022)	0.67	0.63	0.73	0.68	**0.83**	**0.70**	0.58	**0.27**	0.56	0.03	0.96	0.02
Non-uniform	**0.75**	0.62	**0.83**	0.68	0.80	0.48	0.58	0.04	0.67	**0.04**	**1.00**	0.00
SAURON	0.64	**0.64**	0.71	**0.73**	0.78	0.68	0.55	0.25	0.49	**0.04**	0.95	0.09

Table 4.12: Table comparing the performance of the RoBERTa base model, the model from Wegmann, Schraagen and Nguyen (2022), the previously proposed model ("Non-uniform"), and the new proposed model ("SAURON") on the AV and CAV tasks using 0% and 100% semantically similar different-author examples, as well as the STEL framework. The table shows the results for both the original tasks and the STEL-or-content (o-c) tasks. The values with the highest accuracy in each column are reported in bold.

The results, as shown in Table 4.12, reveal several intriguing points about the SAURON model. This new approach heavily outperforms the previous model on the STEL-or-content tasks. Particularly, the model's performance on the formal dimension jumped by 20 percentage points, while it increased by 21 points on the complex dimension. For the contraction, this model even scores the highest among all other style embedding models (see Table 4.3), excluding the RoBERTa base model. While this performance is still not superior to the one achieved by Wegmann, Schraagen and Nguyen (2022), it does validate the assertion that incorporating a form of content control can yield improved style representation. The reasons why it doesn't surpass the performance of Wegmann, Schraagen and Nguyen (2022) align with the potential pitfalls identified in the previous sections.

Interestingly, the SAURON model outshines all the other models on the 100% AV and CAV tasks, indicating the effectiveness of the new approach in style-content disentanglement when semantically similar utterances are considered. However, when this condition is removed (0% tasks), the model's relative performance drops considerably, landing the worst results among all models. Despite this, it is noteworthy that the

accuracies that the model achieved in the 0% and 100% tasks are reasonably close, suggesting that the model's capability of distinguishing authors remains relatively unchanged regardless of the proportion of semantically similar utterances. This implies a certain robustness of the SAURON model in handling variations in the semantic content of the utterances, thereby maintaining its performance in Authorship Verification tasks even when the degree of semantic similarity is altered.

Regarding the original STEL tasks, the SAURON model performs worse than the "Non-uniform" model and the model from Wegmann, Schraagen and Nguyen (2022). This indicates that there is still room for improvement, particularly in designing models that can excel in both the STEL tasks and the style-content disentanglement tasks. However, this experiment confirms the promising potential of integrating the Triplet loss function and other methodological modifications inspired by Wegmann, Schraagen and Nguyen (2022) in the task of style-content disentanglement.

Of notable importance is that this model was developed without any hyperparameter tuning for the loss function or optimization of the "margin" parameter in the Triplet loss. The potential for enhanced performance through such adjustments is an area that could be fruitfully explored in future work.

4.9 MAIN DISCUSSION

In this main discussion section, I undertake a comprehensive comparison of all the models developed throughout the course of this thesis. I examine their performances and individual characteristics, highlighting noteworthy differences and commonalities. The aim of this discussion is not only to understand the strengths and weaknesses of each model in isolation but also to draw meaningful conclusions about the broader implications of different modeling strategies for capturing writing style. This section will provide an overarching narrative, linking the various aspects of the research and drawing out key points for discussion. It is intended to offer a clear synthesis of the experimental findings, grounding them within the broader context of the research question, and outlining potential pathways for future exploration in this field.

4.9.1 *(Contrastive) Authorship Verification*

Overall, the approach employed in this study exhibited masterful performance in the Authorship Verification and Contrastive Authorship Verification tasks. In Table 4.13 we can see that my approach systematically outperforms the base RoBERTa model, as well as improve or perform similarly to the model from Wegmann, Schraagen and Nguyen (2022) on both the AV and CAV tasks. The only exception to this trend is the "Positive-negative" model, which all style representation models on the 100% tasks outperform.

	AV task		CAV task		Formal		Complex		Nb3r		C'tion	
Model	0%	100%	0%	100%	Original	o-c	Original	o-c	Original	o-c	Original	o-c
RoBERTa	0.58	0.50	0.63	0.57	**0.83**	0.09	**0.73**	0.01	**1.00**	0.00	0.94	**0.13**
Wegmann et al. (2022)	0.67	0.63	0.73	0.68	**0.83**	**0.70**	0.58	**0.27**	0.56	0.03	0.96	0.02
Initial model	0.72	0.60	0.80	0.64	0.81	0.48	0.59	0.04	0.65	**0.06**	0.95	0.00
Adjusted authors	0.75	0.62	0.82	0.68	0.78	0.47	0.57	0.04	0.72	0.03	**1.00**	0.00
Positive-negative	**0.78**	0.55	**0.85**	0.55	0.76	0.16	0.58	0.00	0.63	0.04	**1.00**	0.00
Random sampling	0.73	0.61	0.83	0.67	0.81	0.48	0.55	0.05	0.61	0.03	**1.00**	0.00
Non-uniform	0.75	0.62	0.83	0.68	0.80	0.48	0.58	0.04	0.67	0.04	**1.00**	0.00
SAURON	0.64	**0.64**	0.71	**0.73**	0.78	0.68	0.55	0.25	0.49	0.04	0.95	0.09

Table 4.13: Table comparing the performance of the RoBERTa base model, the model from Wegmann, Schraagen and Nguyen (2022), and all models presented in this thesis on the AV and CAV tasks using 0% and 100% semantically similar negative examples, as well as the STEL framework. The Formal, Complex, Number substitution and Contraction task columns are subdivided into Original (standard STEL dataset) and o-c (STEL-or-content dataset) subcolumns. The values with the highest accuracy in each column are reported in **bold**. The underlined values correspond to the highest accuracy in each column for the models that I present throughout this chapter. **Note:** This table is the same as Table 4.3.

Most interestingly, the SAURON model, which introduced a setup that more closely resembled Wegmann, Schraagen and Nguyen (2022), emerged with the best performance on the 100% tasks, surpassing all previous models. These results provide evidence supporting the efficacy of the SAURON model in situations with a high proportion of semantically similar different-author examples, attesting to its ability to balance the intricate interaction between style disentanglement and content control.
The relatively consistent performance of the SAURON model in both the 0% and 100% tasks presents a noteworthy observation. Despite having the lowest score in the 0% tasks, the closeness of the scores between these two extremes suggests that the model's capacity to differentiate authors remains relatively stable, regardless of the proportion of semantically similar different-author examples. This result indicates the model's potential robustness in Authorship Verification tasks.

Conversely, the "Positive-negative" model, as discussed in its respective subsection (§ 4.5), demonstrates superior performance in the 0% tasks, where no semantically similar different-author examples are used. It excels in both the AV and CAV tasks, reaching an accuracy of 0.78 and 0.85, respectively. However, as the proportion of semantically similar different-author examples increases to 100%, the model's performance diminishes enormously. This phenomenon underscores the complex balance that needs to be maintained between ensuring content control and achieving effective style disentanglement.

4.9.2 *STEL tasks*

Continuing from my discussion on the "Positive-negative" model in the context of the AV and CAV tasks, this model's inability to discern style from content becomes more

pronounced in the STEL-or-content tasks. Despite demonstrating superior performance in "easy" AV and CAV tasks, it struggles to perform effectively in the STEL-or-content tasks. As seen in Table 4.13, it underperforms substantially in these tasks, reaching a maximum accuracy of 0.16 on the formal/informal dimension (vs. 0.70 from Wegmann, Schraagen and Nguyen (2022)), and 0.00 on the simple/complex dimension (vs. 0.27) — a result starkly contrasting its intended purpose.

Regarding the remaining models that were trained using the *MultipleNegativesRankingLoss* (i.e., "Initial model", "Adjusted authors", "Random sampling", "Non-uniform"), similar trends are observed in the STEL-or-content tasks, with only marginal gains between iterations. For the formal/informal and simple/complex dimensions, none of the models substantially outperform the others, and all fell short when compared to the model from Wegmann, Schraagen and Nguyen (2022). This performance disparity suggests potential improvements in the training procedure that could be harnessed to achieve better results in these tasks.

In contrast to its predecessors, the SAURON model managed to attain considerably improved results in these tasks. While earlier models barely exceeded 0.48 accuracy in the formal/informal STEL-or-content task, the SAURON model almost reached the performance level of Wegmann, Schraagen and Nguyen (2022), hitting 0.68 accuracy. Similarly, in the simple/complex STEL-or-content task, the SAURON model achieved an accuracy of 0.25, a huge leap from the maximum 0.05 observed in earlier models. However, it's worth noting that despite these improvements, the SAURON model still fell short of the performance of Wegmann, Schraagen and Nguyen (2022), suggesting there are further enhancements to be made in the models' approach to the STEL-or-content tasks. One potential direction for improvement could be conducting a more extensive hyperparameter search specifically tailored for these tasks, to optimize the model's ability to accurately discern style from content. Also, exploring the use of different initial seeds during training might provide some performance variance, offering yet another avenue for potential improvements.

The results from the original STEL tasks offer some interesting insights. Although they serve as an initial benchmark, the observed performance does not correlate strongly with the more complex tasks, such as the STEL-or-content tasks. For instance, the "Positive-negative" model, despite its apparent poor ability to unravel style from content, shows competitive performance on the simple/complex, number substitution, and contraction dimensions. This observation raises questions about the efficacy of the original STEL tasks as an indicative measure for evaluating models designed for more nuanced style-content disentanglement. The disparity suggests that success in these original tasks does not necessarily translate to similar proficiency in more challenging tasks. Consequently, the argument could be made that these tasks might be less useful

for informing the development of models intended for more complex tasks involving style-content disentanglement.

4.9.3 *Conclusion of proposed approaches*

In examining the performance of the various approaches used in this thesis, it remains clear that no single model conclusively outperforms the rest across every task. Each model exhibits its own strengths and weaknesses, demonstrating prowess in certain tasks while lagging in others. For instance, while the "Positive-negative" model shows superior performance in the AV and CAV tasks, it falters in the more intricate STEL-or-content tasks, struggling to effectively separate style from content. On the other hand, models like "Adjusted authors", "Random sampling", and the previous contender "Non-uniform", display relative consistency in their performance across all tasks.

That being said, the SAURON model makes a strong case for being considered the most promising solution. Although it doesn't achieve the highest scores in the 0% AV and CAV tasks, it is the strongest performer in the 100% AV and CAV tasks and outperforms all models in the STEL-or-content tasks. Moreover, its performance on the AV and CAV tasks is fairly robust. The SAURON model successfully incorporates benefits observed in previous models, further fortifying its capabilities. Building on the "Adjusted authors" model, it enhances the breadth of available data by incorporating one utterance authors. The addition of the random selection of utterances - rather than using the first 10 chronological - from the "Random sampling" model mitigates the potential for overrepresentation and bias arising from authors having too many utterances from the same conversation. Importantly, it introduces a new advantage: it allows for a more varied set of paraphrases as different-author examples. This approach not only enhances the model's robustness but also makes it more adept at handling real-world data, where stylistic expressions can be quite diverse.

Nonetheless, it is worth noting that the model from Wegmann, Schraagen and Nguyen (2022) still outperforms all my models in most of the STEL tasks. This stands as an important reminder of the room for further improvement and refinement in my training and modeling techniques.
While it might be too early to definitively label the SAURON model as the "best", its performance and the innovative strategies it incorporates undoubtedly showcase its potential. With its capability to incorporate the strengths of different approaches, the SAURON model signifies a substantial stride toward separating style from content in complex AV/CAV and STEL tasks. Looking ahead, it will be crucial to build upon these findings, focusing on strategies that can further enhance my model's ability to manage increasing data complexity and scale.

5

CONCLUSION

This chapter brings together the various threads of the research presented in this thesis, providing a comprehensive summary of the findings and reflecting on their implications. The aim is to revisit the research questions outlined in Section § 1.3, providing clear and concise answers based on the evidence gathered throughout the course of this study. Following the summary of findings, this chapter will delve into a discussion of the limitations of the current study. Recognizing these limitations not only ensures a balanced and honest reflection on the work conducted but also helps to identify areas where further research could be beneficial.

5.1 RESEARCH QUESTION

How does incorporating semantically similar utterances affect the performance of transformer-based approaches for writing style representation? To comprehensively answer this primary research question, I will delve into two related subquestions: first, examining the effect of using semantically similar utterances in comparison with other control data; and second, investigating the impact of various sampling methods for these utterances. Analyzing these facets will provide a thorough understanding to address my main research question.

5.1.1 *Subquestion 1*

When addressing the first research subquestion, *"How does the use of semantically similar utterances compare to using other types of control data (e.g. unrelated sentences, sentences from the same conversation)?"*, the data clearly suggests that the use of semantically similar utterances can indeed positively impact the model's ability to disentangle style from content. As seen from the results of the initial experiments, the introduction of semantically similar different-author examples led to marked improvements in both the STEL and AV/CAV tasks (Section § 4.3.2). These improvements were observed in comparison to the models that either did not use semantically similar different-author examples at all or only incorporated them partially. To illustrate, the accuracy for the formal/informal STEL-or-content dimension improved systematically, from 0.24 to 0.48, as the proportion of semantically similar different-author examples was increased from 0% to 100%. Similarly, the accuracy on the 100% AV and CAV tasks increased from 0.53 to 0.60 and from 0.53 to 0.64, respectively, when increasing the ratio of semantically similar

different-author examples.

The pivotal shift in this research comes with the introduction of the SAURON model. This model, which utilizes semantically similar utterances and innovatively adjusts the loss function and use of anchors, substantially improved upon the earlier models. In fact, it outperformed both the initial models and the model from Wegmann, Schraagen and Nguyen (2022) where different-author examples are sampled from the same topic (i.e., subreddit), as well as their model that does not use any content control at all. The SAURON model does not only achieve better performance on the 100% AV and CAV tasks but also surpasses all models in the STEL-or-content tasks, cementing the effectiveness of the semantically similar utterance approach. However, it is important to note that the SAURON model still underperforms the conversation-level control model from Wegmann, Schraagen and Nguyen (2022) in several STEL tasks, indicating that there remains room for further improvement.

It can thus be concluded that despite these advancements, the semantically similar approach still exhibits some limitations, especially when compared to the conversation-based sampling method. The restrictions on the length of the paraphrases and the tendency to favor high-frequency paraphrases in the semantically similar approach can limit their diversity. Conversely, the conversation-based approach potentially exposes the model to a wider range of stylistic expressions, which may account for its better performance in some STEL tasks.

These observations provide valuable directions for future research. It is plausible that an optimal balance might be struck by integrating the stylistic diversity benefits gleaned from conversation-based sampling with the semantic similarity aspect, which could potentially enhance the model's capacity to effectively disentangle style from content. Alternatively, future advancements could also be achieved by further refining the semantic similarity approach itself.

5.1.2 *Subquestion 2*

The results of the experiments reveal several key strategies that can be beneficial for answering the subquestion *"What are the most effective sampling techniques for preparing the input data?"* The following aspects emerged as crucial considerations:

- **Author diversity:** It is essential to incorporate as many different authors as possible, as demonstrated in § 4.4. By doing this, the model gets exposure to a vast range of distinct writing styles, which enhances its ability to recognize and differentiate between various styles.
- **Balanced representation:** Section § 4.4 also showed that the overrepresentation of certain authors should be avoided. In this study, the maximum number of utterances per author was set to 10. However, this value should be adjusted based on the size

and diversity of one's dataset. Balanced representation helps the model avoid developing a bias towards overrepresented authors.

- **Topic Diversity:** Broadening the array of topics in the dataset is instrumental for enhancing the model's capacity to generalize across diverse topics, as discussed in the introduction of § 4.4. This approach ensures the model does not cultivate a bias towards certain topics. Even though the available dataset constrained the number of conversations, the experiment in § 4.6 did highlight the marginal benefits of expanding topic diversity. However, the potential value of increasing the topic diversity specifically within each author's work was less explored. The theory here is that the richness of an author's writing style could be further exposed by incorporating varied topics from them, as an author's style may subtly fluctuate depending on the topic. Yet, the dataset's limited size and scope might have precluded a comprehensive investigation into this aspect in the experiment of Section § 4.6. Therefore, while the marginal benefits of more diverse topics are apparent, future research with larger, more diverse datasets might be necessary to concretely affirm the advantages of within-author topic diversity.

- **Paraphrase diversification:** The thesis also highlights the need to avoid oversampling the same top paraphrases. A diverse set of paraphrases leads to a more robust and nuanced understanding of style, as it prevents the model from overfitting to particular phrases or structures — something that is expanded upon in Section § 4.7.

5.1.3 *Answer to main research question*

The results of this thesis help answer the main research question: *"How does incorporating semantically similar utterances affect the performance of transformer-based approaches for writing style representation?"* By investigating various methodologies and conducting extensive experiments, it has been established that integrating semantically similar utterances does indeed lead to substantial improvements over not using any form of content control at all. However, it has also become clear that relying exclusively on semantically similar utterances as different-author examples is not the most optimal strategy.

The use of semantically similar utterances as different-author examples, as detailed in subquestion 1, led to substantial improvements in both the STEL and AV/CAV tasks. Particularly, these gains were amplified when the model, which was previously trained on data devoid of these utterances, was subsequently trained with an increased proportion of semantically similar utterances. This underlines the effectiveness of incorporating semantically similar utterances. However, it's important to note that this method fell short of the approach where different-author examples are sampled from the same conversation, as per the methodology of Wegmann, Schraagen and Nguyen (2022) — their method achieved better performance in almost all STEL tasks and comparable performance in the 100% AV and CAV tasks.

The disparities in results between the two methodologies suggest that there might be

a potential sweet spot in combining the two approaches. This could involve integrating semantically similar utterances, while also allowing for a less strict sampling approach, which would permit more diverse examples.

In response to subquestion 2, the effectiveness of input data preparation techniques was also examined. The key takeaways include the importance of incorporating as many different authors as possible to increase the number of distinct writing styles, ensuring topic diversity within each author and across the dataset, avoiding overrepresentation of the same authors, and diversifying the sampling of top paraphrases.

In conclusion, while incorporating semantically similar utterances does enhance the performance of transformer-based models in writing style representation tasks, it should not be viewed as the sole solution. Future research should look into combining this approach with others, such as the conversation-based sampling of different-author examples proposed by Wegmann, Schraagen and Nguyen (2022), and further optimizing the input data preparation techniques. Such strategies can potentially lead to more robust and nuanced style representations and improved performance in various tasks.

5.2 LIMITATIONS AND FUTURE WORK

The development and evaluation of writing style embedding models, such as the one proposed in this thesis, are subject to several limitations that should be acknowledged. The forthcoming subsections aim to offer a comprehensive overview of these limitations. They cover a spectrum of constraints, from methodological to data-based to evaluation-related, each contributing to the multifaceted challenge of this research area. This collection is not organized hierarchically but is intended to present a realistic picture of the constraints encountered in this study.

5.2.1 *Evaluation methods*

While there is a vast amount of data available, the field of style modeling lacks a standardized benchmark dataset. Without a standard benchmark dataset, determining the quality of a style representation model becomes a challenging task. Although the STEL framework from Wegmann and Nguyen (2021) is a step in the right direction, their framework also has its limitations.
For instance, the limited number of examples for each dimension in the dataset ($n = 815$ for the formal and complex dimensions, and $n = 100$ for number substitution and contraction) can impact the reliability of the results. This scarcity of examples also limits the ability to test for small, subtle differences in style, which could be crucial for certain applications. Moreover, the STEL framework may not be fully representative of the stylistic nuances present in the data. This lack of representativeness could cause models to not exhibit significant differences in performance, thereby limiting the usefulness

of the framework for evaluating the style-measuring capability of different models. Finally, while the STEL framework uses both complex style dimensions and simpler characteristics, such as contraction and number substitution, it may not capture all the relevant dimensions of writing style. This limited scope can impact the ability to fully evaluate the style-measuring capability of different models and suggests a need for more comprehensive frameworks in future research.

Future work could focus on expanding the STEL framework or creating a different benchmark, perhaps through the use of crowd-sourcing or other data collection methods. This could help to create a more robust and representative benchmark dataset for evaluating style representation models.

5.2.2 *Interpretability*

Transformer-based models, despite their impressive performance in various NLP tasks, are often criticized for their lack of interpretability (Chefer, Gur and Wolf, 2020). The complexity of these models, coupled with their multi-layered architecture and millions of parameters, makes it difficult to understand their internal workings and to determine why a particular prediction was made. This opacity, often referred to as the 'black box' problem, can limit their applicability in scenarios where the reasoning behind a prediction is crucial. In the context of style representation, understanding why a model identifies two pieces of text as having a similar style could provide valuable insights into the characteristics of writing style that the model is capturing.
Moreover, the lack of interpretability can also hinder the process of model improvement. Without a clear understanding of how the model is making its decisions, it is difficult to identify the sources of errors or areas where the model's performance could be improved. This can slow down the process of model refinement and limit the potential for performance gains.

Future work in this area could focus on developing methods to improve the interpretability of transformer-based models, specifically for style representation. This could involve techniques such as attention visualization (Vig, 2019), or the development of inherently interpretable models. While some of these techniques are available for classification tasks, they are still in the early stages of representation learning. By improving the interpretability of these models, we can not only enhance their usefulness in various applications but also gain deeper insights into the nature of writing style itself.

5.2.3 *The definition of style*

The lack of a universally agreed-upon definition of writing style complicates the development and evaluation of writing style embedding models. Without clear criteria for what constitutes a good style representation, the task of developing an effective model becomes

more challenging.
In this thesis, a definition for style is provided in Section § 2.1. However, it's important to note that this definition is not widely accepted across the field. Different researchers may have different interpretations of what constitutes writing style, and different models may focus on different aspects of style. This lack of consensus can complicate the task of developing and evaluating writing style embedding models. Without a clear target to aim for, it becomes more challenging to determine whether a model is effectively capturing the nuances of writing style.

5.2.4 *Semantic optimization*

The incorporation of semantically similar utterances improves the performance of writing style embedding models over models that do not employ any content control at all but is still outperformed by models that employ conversation-level content control. Determining the optimal parameters for generating these utterances is a non-trivial task. The effectiveness of the proposed approach can be impacted by these choices, and finding the right balance between the quality and the number of paraphrases requires careful consideration and experimentation. For instance, future work could explore how the size of the dataset from which paraphrases are sampled impacts their quality, as a larger dataset potentially offers a broader selection of suitable matches.
Moreover, the impact of scores on the quality of the paraphrases is another aspect that warrants further investigation. While higher scores generally indicate greater semantic similarity, the optimal score threshold for generating high-quality paraphrases may not be immediately apparent and could vary depending on the specific task or application. The analysis of the sampled paraphrases in Section § 4.1 shows that paraphrases with a lower similarity score can also show similar content. The experimentation with the similarity threshold is also something that future work can explore.

Additionally, two limiting factors were also touched upon in Section § 4.1.2: the frequency distribution showed that (i) a limited number of paraphrases were disproportionally represented in the dataset and (ii) the dataset predominantly consisted of shorter, similar paraphrases. The problem here is not necessarily with the paraphrase mining approach I utilized, but rather with the fact that - in general - it is easier to find semantically similar utterances for shorter sentences, and that the Reddit data source also contains a lot of these short utterances.

As a possible solution for the issues raised in this subsection, the approach proposed in this thesis could potentially be combined with other methods, such as the one proposed by Wegmann, Schraagen and Nguyen (2022). For example, one could first attempt to generate as many semantically similar examples as possible using the approach proposed in this thesis, and then for the remaining texts, sample from within the same conversation as per their method. This hybrid approach could potentially leverage the

strengths of both methods and is another promising direction for future work.
Another possible avenue to explore could be to split the utterances into sentences to take advantage of the fact that paraphrases are easier to find for shorter texts. This strategy could substantially increase the number of paraphrases available, thereby augmenting the dataset. However, potential drawbacks of this approach might be that the model becomes less adept at handling longer texts, and some stylistic nuances could be lost in the process of breaking down longer utterances into shorter sentences.

5.2.5 *Data diversity*

The scope of this work is confined to the context of data from Reddit, specifically, a selected sample of 100 subreddits. While this platform indeed provides a rich source of different writing styles, it is important to bear in mind that the findings of this study are contingent upon this specific data source and the limited number of subreddits considered. Therefore, the effectiveness of the methodologies used might not seamlessly generalize to other types of written language, different social media platforms, or even other subreddits outside the ones used in this study. This limitation could restrict the applicability of the approach to certain applications.

Further expanding the diversity of data sources could be a promising direction for future research. As was briefly discussed in the introduction of this thesis (Chapter 1), different social media platforms such as Twitter, Facebook, and even distinct subreddits within Reddit, each host unique writing conventions and styles (Marko and Buker, 2022). Therefore, integrating data from a wider array of sources, or considering a more diverse selection of subreddits, could enhance the model's understanding of various writing styles, and potentially improve its capacity to disentangle style from content. Ultimately, diversifying data sources could broaden the applicability of the techniques developed in this study, potentially offering more robust and generalized models of authorial style.

5.2.6 *Time constraints*

This study, while comprehensive in its approach, was subject to certain time constraints that limited the extent to which the research could be conducted. Some of these limitations are outlined below.

Firstly, only one seed was used for initializing each model. As a result, the outcomes of the model training processes can be largely dependent on the choice of this initial seed. The variability in performance among different seeds can lead to unstable results, with some seeds potentially providing better outcomes than others. Although the models used in this study yielded insightful findings, they may not represent the "best" or most optimal models that could have been produced with a different initial seed. Hence, the generalizability of the results may be somewhat limited. Future work should include

more runs with different seeds to average the results and obtain a more reliable estimate of the model's performance.

Secondly, while potential benefits of a larger dataset have been highlighted in previous discussions in Sections § 4.1 and § 4.4, such as providing more diverse and representative samples of various writing styles, the decision was made not to pursue this expansion. Increasing the dataset's size would have required additional computational resources and processing time. Given the scope of this study and the need to focus on methodological developments, the decision was taken to work with the current dataset size. While this decision streamlined the project and allowed for a concentrated focus on the techniques being developed, it may have had implications for the scope and diversity of writing styles the models could learn and represent.

Finally, the SAURON model implemented the approach of using each anchor only once, following the method used by Wegmann, Schraagen and Nguyen (2022). This was a valuable development in reducing the overrepresentation of certain phrases. However, due to time constraints, the effects of using each anchor multiple times, as was done in previous models, were not explored within this setup. Additionally, potential performance improvements from tuning hyperparameters in the SAURON model setup remain unexplored. Future work could look into these aspects to possibly enhance the model's performance and robustness. The implementation and evaluation of these methods would, however, require additional time and computational resources.

A

HYPERPARAMETER TUNING

A.1 LOSS FUNCTION

	AV task		CAV task		Formal		Complex		Nb3r		C'tion	
Loss function	0%	100%	0%	100%	Original	o-c	Original	o-c	Original	o-c	Original	o-c
MNRL	**0.72**	**0.60**	**0.80**	**0.64**	**0.81**	**0.48**	**0.59**	**0.04**	**0.65**	0.06	**0.95**	**0.00**
Contrastive	0.70	0.52	0.78	0.53	0.78	0.42	0.56	**0.04**	0.62	**0.07**	0.93	**0.00**

Table A.1: The table above presents the results of the impact of two different loss functions on the test tasks. Columns for the AV Task and CAV Task show results for 0% and 100% semantically similar conditions, respectively. The Formal, Complex, Number substitution, and Contraction task columns are subdivided into Original (standard STEL task) and o-c (STEL-or-content task) subcolumns. The values with the highest accuracy in each column are reported in bold.

A.2 LEARNING RATE

	AV task		CAV task		Formal		Complex		Nb3r		C'tion	
Learning rate	0%	100%	0%	100%	Original	o-c	Original	o-c	Original	o-c	Original	o-c
1e-5	**0.72**	0.59	**0.80**	0.62	**0.82**	0.47	**0.59**	**0.04**	0.64	0.05	0.97	0.00
2e-5	**0.72**	**0.60**	**0.80**	**0.64**	0.81	0.48	**0.59**	**0.04**	0.65	**0.06**	0.95	0.00
3e-5	**0.72**	0.56	**0.80**	0.63	0.79	**0.49**	0.58	**0.04**	0.63	0.04	**1.00**	0.00
4e-5	**0.72**	0.59	**0.80**	0.62	0.79	0.46	0.55	**0.04**	**0.71**	0.00	0.96	**0.08**

Table A.2: The table above presents the results of the impact of various learning rates on the test tasks. Columns for the AV Task and CAV Task show results for 0% and 100% semantically similar conditions, respectively. The Formal, Complex, Number substitution, and Contraction task columns are subdivided into Original (standard STEL task) and o-c (STEL-or-content task) subcolumns. The values with the highest accuracy in each column are reported in bold.

A.3 NUMBER OF EPOCHS

	AV task		CAV task		Formal		Complex		Nb3r		C'tion	
Epochs	0%	100%	0%	100%	Original	o-c	Original	o-c	Original	o-c	Original	o-c
2	0.70	0.56	0.80	0.62	0.77	0.24	0.57	0.02	0.58	0.04	0.92	**0.01**
3	0.70	0.57	**0.81**	0.62	0.78	0.34	**0.60**	0.04	0.63	0.04	**0.99**	**0.01**
4	0.71	0.59	0.80	0.63	0.80	0.40	0.55	0.03	0.64	0.03	0.96	0.00
5	**0.72**	**0.60**	0.80	**0.64**	**0.81**	**0.48**	0.59	0.04	**0.65**	**0.06**	0.95	0.00
6	**0.72**	**0.60**	0.80	0.64	**0.81**	0.47	0.58	0.04	**0.65**	**0.06**	0.94	0.00

Table A.3: The table above presents the results of the impact of the number of training epochs on the test tasks. Columns for the AV Task and CAV Task show results for 0% and 100% semantically similar conditions, respectively. The Formal, Complex, Number substitution, and Contraction task columns are subdivided into Original (standard STEL task) and o-c (STEL-or-content task) subcolumns. The values with the highest accuracy in each column are reported in bold.

A.4 BATCH SIZE

	AV task		CAV task		Formal		Complex		Nb3r		C'tion	
Batch size	0%	100%	0%	100%	Original	o-c	Original	o-c	Original	o-c	Original	o-c
2	0.69	0.53	0.77	0.60	0.78	0.46	0.57	**0.04**	0.64	0.05	**1.00**	0.00
4	0.70	0.54	0.78	0.63	0.78	0.48	0.56	**0.04**	0.66	0.05	**1.00**	0.00
8	**0.72**	**0.60**	**0.80**	**0.64**	**0.81**	**0.48**	**0.59**	**0.04**	**0.65**	**0.06**	0.95	0.00

Table A.4: The table above presents the results of the impact of various batch sizes. Columns for the AV Task and CAV Task show results for 0% and 100% semantically similar conditions, respectively. The Formal, Complex, Number substitution, and Contraction task columns are subdivided into Original (standard STEL task) and o-c (STEL-or-content task) subcolumns. The values with the highest accuracy in each column are reported in bold.

BIBLIOGRAPHY

Adair, Douglass (1944). 'The Authorship of the Disputed Federalist Papers'. In: *The William and Mary Quarterly* 1.2, pp. 98–122. ISSN: 00435597, 19337698. URL: http://www.jstor.org/stable/1921883 (visited on 16/01/2023).

Addison, Joseph and Richard Steele (1711). 'The Spectator'. In: *The Spectator* 1.1, pp. 1–16.

Amir, Silvio, Byron C. Wallace, Hao Lyu, Paula Carvalho and Mário J. Silva (2016). 'Modelling Context with User Embeddings for Sarcasm Detection in Social Media'. In: *CoRR* abs/1607.00976. arXiv: 1607.00976. URL: http://arxiv.org/abs/1607.00976.

Argamon, Shlomo, Moshe Koppel, Jonathan Fine and Anat Rachel Shimoni (2003). 'Gender, Genre, and Writing Style in Formal Written Texts'. In: *Text & talk* 23.3, pp. 321–346. DOI: doi:10.1515/text.2003.014. URL: https://doi.org/10.1515/text.2003.014.

Argamon, Shlomo and Shlomo Levitan (2005). 'Measuring the Usefulness of Function Words For Authorship Attribution'. In: *Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*, pp. 1–3.

Bergs, Alexander (2015). 'Linguistic Fingerprints of Authors and Scribes'. In: *Letter Writing and Language Change*. Ed. by Anita Auer, Daniel Schreier and Richard J.Editors Watts. Studies in English Language. Cambridge University Press, 114–132. DOI: 10.1017/CBO9781139088275.008.

Biber, Douglas and Susan Conrad (2009). *Register, Genre, and Style*. Cambridge Textbooks in Linguistics. Cambridge University Press. DOI: 10.1017/CBO9780511814358.

Bischoff, Sebastian, Niklas Deckers, Marcel Schliebs, Ben Thies, Matthias Hagen, Efstathios Stamatatos, Benno Stein and Martin Potthast (2020). *The Importance of Suppressing Domain Style in Authorship Analysis*. DOI: 10.48550/ARXIV.2005.14714. URL: https://arxiv.org/abs/2005.14714.

Brooks, Cleanth (1947). *The Well-Wrought Urn: Studies in the Structure of Poetry*.

Buda, Jakab and Flóra Bolonyai (2020). 'An Ensemble Model Using N-grams and Statistical Features to Identify Fake News Spreaders on Twitter Notebook for PAN at CLEF 2020'. In.

Burrows, John (2002). ''Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship'. In: *Literary and Linguistic Computing* 17.3, pp. 267–287. ISSN: 0268-1145. DOI: 10.1093/llc/17.3.267. eprint: https://academic.oup.com/dsh/article-pdf/17/3/267/2743069/170267.pdf. URL: https://doi.org/10.1093/llc/17.3.267.

Chan, C-S (1994). 'Operational Definitions of Style'. In: *Environment and Planning B: Planning and Design* 21.2, pp. 223–246. DOI: 10.1068/b210223. eprint: https://doi.org/10.1068/b210223. URL: https://doi.org/10.1068/b210223.

Chang, Jonathan P., Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang and Cristian Danescu-Niculescu-Mizil (2020). 'ConvoKit: A Toolkit for the Analysis of Conversations'. In: *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 1st virtual meeting: Association for Computational Linguistics, pp. 57–60. URL: https://aclanthology.org/2020.sigdial-1.8.

Chartprasert, Duangkamol (1993). 'How Bureaucratic Writing Style Affects Source Credibility'. In: *Journalism Quarterly* 70.1, pp. 150–159. DOI: 10.1177/107769909307000117. eprint: https://doi.org/10.1177/107769909307000117. URL: https://doi.org/10.1177/107769909307000117.

Chaski, Carole E (2005). 'Who's at the Keyboard? Authorship Attribution in Digital Evidence Investigations'. In: *International journal of digital evidence* 4.1, pp. 1–13.

Chefer, Hila, Shir Gur and Lior Wolf (2020). 'Transformer Interpretability Beyond Attention Visualization'. In: *CoRR* abs/2012.09838. arXiv: 2012.09838. URL: https://arxiv.org/abs/2012.09838.

Cho, KyungHyun, Bart van Merrienboer, Dzmitry Bahdanau and Yoshua Bengio (2014). 'On the Properties of Neural Machine Translation: Encoder-Decoder Approaches'. In: *CoRR* abs/1409.1259. arXiv: 1409.1259. URL: http://arxiv.org/abs/1409.1259.

Choi, Sujin and Jihoon Lim (2019). 'Determinant and Consequence of Online News Authorship Verification: Blind News Consumption Creates Press Credibility'. In: *International Journal of Communication* 13.0. ISSN: 1932-8036. URL: https://ijoc.org/index.php/ijoc/article/view/9594.

Clement, J. (2022). *Topic: Internet Usage Worldwide*. URL: https://www.statista.com/topics/1145/internet-usage-worldwide/.

Coulthard, Malcolm (2014). 'Whose text is it? On the linguistic investigation of authorship'. In: pp. 270–287. ISBN: 9781315838502. DOI: 10.4324/9781315838502-15.

Damerau, Fred J. (1975). 'The Use of Function Word Frequencies as Indicators of Style'. In: *Computers and the Humanities* 9.6, pp. 271–280. ISSN: 00104817. URL: http://www.jstor.org/stable/30204237 (visited on 09/01/2023).

Damerau, Frederick J and Benoît B Mandelbrot (1973). 'Tests of the Degree of Word Clustering in Samples of Written English'. In: 11.102, pp. 58–75. DOI: doi:10.1515/ling.1973.11.102.58.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2019). 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

Dictionary, Cambridge (2023). *Style*. URL: https://dictionary.cambridge.org/dictionary/english/style.

Emerson, Rob (2022). *What does "/S" mean on reddit?* URL: https://www.itgeared.com/what-does-s-mean-on-reddit/.

Fabien, Maël, Esau Villatoro-Tello, Petr Motlicek and Shantipriya Parida (2020). 'BertAA : BERT Fine-Tuning for Authorship Attribution'. In: *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*. Indian Institute of Technology Patna, Patna, India: NLP Association of India (NLPAI), pp. 127–137. URL: https://aclanthology.org/2020.icon-main.16.

Filik, Ruth, Christian Mark Hunter and Hartmut Leuthold (2015). 'When Language Gets Eemotional: Irony and the Embodiment of Affect in Discourse'. In: *Acta Psychologica* 156, pp. 114–125. ISSN: 0001-6918. DOI: https://doi.org/10.1016/j.actpsy.2014.08.007. URL: https://www.sciencedirect.com/science/article/pii/S0001691814001917.

Fraser, Murray (2018). 'Gothic'. In: *Sir Banister Fletcher Glossary*. DOI: 10.5040/9781350122741.1001019.

Funkhouser, G. Ray and Nathan Maccoby (1973). 'Tailoring Science Writing to the General Audience'. In: *Journalism Quarterly* 50.2, pp. 220–226. DOI: 10.1177/107769907305000202. eprint: https://doi.org/10.1177/107769907305000202. URL: https://doi.org/10.1177/107769907305000202.

Gao, Tianyu, Xingcheng Yao and Danqi Chen (Nov. 2021). 'SimCSE: Simple Contrastive Learning of Sentence Embeddings'. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 6894–6910. DOI: 10.18653/v1/2021.emnlp-main.552. URL: https://aclanthology.org/2021.emnlp-main.552.

Goodfellow, Ian J., Jonathon Shlens and Christian Szegedy (2015). 'Explaining and Harnessing Adversarial Examples'. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. URL: http://arxiv.org/abs/1412.6572.

Hacker, Diana (1994). *The Bedford Handbook for Writers*. Bedford Books of St. Martin's Press. ISBN: 9781457683039.

Hadsell, R., S. Chopra and Y. LeCun (2006). 'Dimensionality Reduction by Learning an Invariant Mapping'. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 2, pp. 1735–1742. DOI: 10.1109/CVPR.2006.100.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). 'Long Short-Term Memory'. In: *Neural Computation* 9.8, 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

Holmes, David (1994). 'Authorship Attribution'. In: *Computers and the Humanities* 28.2, pp. 87–106. ISSN: 00104817. URL: http://www.jstor.org/stable/30200315 (visited on 26/06/2023).

Holmes, David and Richard S Forsyth (1995). 'The Federalist Revisited: New Directions in Authorship Attribution'. In: *Literary and Linguistic Computing* 10.2, pp. 111–127. ISSN: 0268-1145. DOI: 10.1093/llc/10.2.111. eprint: https://academic.oup.com/dsh/article-pdf/10/2/111/10887101/111.pdf. URL: https://doi.org/10.1093/llc/10.2.111.

Hoover, David L (2004). 'Testing Burrows's delta'. In: *Literary and linguistic computing* 19.4, pp. 453–475.

— (2012). 'The Tutor's Story: A Case Study of Mixed Authorship'. In: *English Studies* 93.3, pp. 324–339. DOI: 10.1080/0013838X.2012.668791. eprint: https://doi.org/10.1080/0013838X.2012.668791. URL: https://doi.org/10.1080/0013838X.2012.668791.

Jafariakinabad, Fereshteh, Sansiri Tarnpradab and Kien A. Hua (2019). 'Syntactic Recurrent Neural Network for Authorship Attribution'. In: *CoRR* abs/1902.09723. arXiv: 1902.09723. URL: http://arxiv.org/abs/1902.09723.

Juola, Patrick (2008). 'Authorship Attribution'. In: *Foundations and Trends in Information Retrieval* 1.3, pp. 233–334. ISSN: 1554-0669. DOI: 10.1561/1500000005. URL: http://dx.doi.org/10.1561/1500000005.

Juslin, Patrik and John Sloboda (2001). *Music and emotion, theory and research*. DOI: 10.1037/1528-3542.1.4.381.

Kaplan, Milton A. (1968). 'Style IS Content'. In: *The English Journal* 57.9, pp. 1330–1334. ISSN: 00138274. URL: http://www.jstor.org/stable/812144 (visited on 28/01/2023).

Kent, Thomas (1986). *Interpretation and Genre: The Role of Generic Perception in the Study of Narrative Texts*. Bucknell University Press.

Keshmiri, Fahimeh and Mina Mahdikhani (2015). 'F. Scott Fitzgerald's Unique Literary and Writing Style'. In: *English Language and Literature Studies* 5.2, p. 78. DOI: 10.5539/ells.v5n2p78.

Kestemont, Mike (2014). 'Function Words in Authorship Attribution. From Black Magic to Theory?' In: *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 59–66. DOI: 10.3115/v1/W14-0908. URL: https://aclanthology.org/W14-0908.

Kestemont, Mike, Kim Luyckx, Walter Daelemans and Thomas Crombez (2012). 'Cross-Genre Authorship Verification Using Unmasking'. In: *English Studies* 93.3, pp. 340–356. DOI: 10.1080/0013838X.2012.668793. eprint: https://doi.org/10.1080/0013838X.2012.668793. URL: https://doi.org/10.1080/0013838X.2012.668793.

Kjell, Bradley (1994). 'Authorship Determination Using Letter Pair Frequency Features with Neural Network Classifiers'. In: *Literary and Linguistic Computing* 9.2, pp. 119–124. ISSN: 0268-1145. DOI: 10.1093/llc/9.2.119. eprint: https://academic.oup.com/dsh/article-pdf/9/2/119/10887227/119.pdf. URL: https://doi.org/10.1093/llc/9.2.119.

Kjell, Bradley, W.Addison Woods and Ophir Frieder (1994). 'Discrimination of Authorship Using Visualization'. In: *Information Processing & Management* 30.1, pp. 141–150. ISSN: 0306-4573. DOI: https://doi.org/10.1016/0306-4573(94)90029-9. URL: https://www.sciencedirect.com/science/article/pii/0306457394900299.

Koppel, Moshe, Shlomo Argamon and Anat Rachel Shimoni (2002). 'Automatically Categorizing Written Texts by Author Gender'. In: *Literary and Linguistic Computing* 17.4, pp. 401–412. ISSN: 0268-1145. DOI: 10.1093/llc/17.4.401. eprint: https://academic.oup.com/dsh/article-pdf/17/4/401/3345463/170401.pdf. URL: https://doi.org/10.1093/llc/17.4.401.

Koppel, Moshe and Jonathan Schler (2004). 'Authorship Verification as a One-Class Classification Problem'. In: ICML '04. Banff, Alberta, Canada: Association for Computing Machinery, p. 62. ISBN: 1581138385. DOI: 10.1145/1015330.1015448. URL: https://doi.org/10.1145/1015330.1015448.

Koppel, Moshe, Jonathan Schler and Shlomo Argamon (2009). 'Computational Methods in Authorship Attribution'. In: *Journal of the American Society for Information Science and Technology* 60.1, pp. 9–26. DOI: https://doi.org/10.1002/asi.20961. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.20961. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20961.

Lample, Guillaume, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato and Y-Lan Boureau (2019). 'Multiple-Attribute Text Rewriting'. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=H1g2NhC5KQ.

Liu, Angela Xia, Ying Xie and Jurui Zhang (2019). 'It's Not Just What You Say, But How You Say It: The Effect of Language Style Matching on Perceived Quality of Consumer Reviews'. In: *Journal of Interactive Marketing* 46, pp. 70–86. ISSN: 1094-9968. DOI: https://doi.org/10.1016/j.intmar.2018.11.001. URL: https://www.sciencedirect.com/science/article/pii/S1094996818300689.

Liu, Xiaodong, Kevin Duh, Liyuan Liu and Jianfeng Gao (2020). 'Very Deep Transformers for Neural Machine Translation'. In: *CoRR* abs/2008.07772. arXiv: 2008.07772. URL: https://arxiv.org/abs/2008.07772.

Liu, Yang and Mirella Lapata (2019). 'Text Summarization with Pretrained Encoders'. In: *CoRR* abs/1908.08345. arXiv: 1908.08345. URL: http://arxiv.org/abs/1908.08345.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov (2019). 'RoBERTa: A Robustly Optimized BERT Pretraining Approach'. In: *CoRR* abs/1907.11692. arXiv: 1907.11692. URL: http://arxiv.org/abs/1907.11692.

Ludwig, Stephan, Ko de Ruyter, Mike Friedman, Elisabeth C. Brüggen, Martin Wetzels and Gerard Pfann (2013). 'More than Words: The Influence of Affective Content and Linguistic Style Matches in Online Reviews on Conversion Rates'. In: *Journal of Marketing* 77.1, pp. 87–103. DOI: 10.1509/jm.11.0560. eprint: https://doi.org/10.1509/jm.11.0560. URL: https://doi.org/10.1509/jm.11.0560.

Manolache, Andrei, Florin Brad, Elena Burceanu, Antonio Barbalau, Radu Tudor Ionescu and Marius Popescu (2021). 'Transferring BERT-like Transformers' Knowledge for Authorship Verification'. In: *CoRR* abs/2112.05125. arXiv: 2112.05125. URL: https://arxiv.org/abs/2112.05125.

Marko, Karoline and Grace Sullivan Buker (2022). '"Hope you're in the mood for Cookies": An Exploratory Study of Individual Writing Styles Across Social Media Platforms'. In: *Journal of Indonesian Community for Forensic Linguistics* 1.1, pp. 14–25.

Mosteller, Frederick and David L. Wallace (1963). 'Inference in an Authorship Problem'. In: *Journal of the American Statistical Association* 58.302, 275–309. DOI: 10.1080/01621459.1963.10500849.

Nallapati, Ramesh, Bing Xiang and Bowen Zhou (2016). 'Sequence-to-Sequence RNNs for Text Summarization'. In: *CoRR* abs/1602.06023. arXiv: 1602.06023. URL: http://arxiv.org/abs/1602.06023.

OpenAI (2023). *GPT-4 Technical Report*. arXiv: 2303.08774 [cs.CL].

Ouamour, S. and H. Sayoud (2013). 'Authorship Attribution of Short Historical Arabic Texts Based on Lexical Features'. In: *2013 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pp. 144–147. DOI: 10.1109/CyberC.2013.31.

Paisley, William J. (1964). 'Identifying the Unknown Communicator in Painting, Literature and Music: The Significance of Minor Encoding Habits'. In: *Journal of Communication* 14.4, pp. 219–237. ISSN: 0021-9916. DOI: 10.1111/j.1460-2466.1964.tb02925.x. eprint: https://academic.oup.com/joc/article-pdf/14/4/219/22386618/jjnlcom0219.pdf. URL: https://doi.org/10.1111/j.1460-2466.1964.tb02925.x.

Pennebaker, James W. (2011a). 'The Secret Life of Pronouns'. In: *New Scientist* 211.2828, pp. 42–45. ISSN: 0262-4079. DOI: https://doi.org/10.1016/S0262-4079(11)62167-2. URL: https://www.sciencedirect.com/science/article/pii/S0262407911621672.

Pennebaker, James W. (2011b). 'Using Computer Analyses to Identify Language Style and Aggressive Intent: The Secret Life of Function Words'. In: *Dynamics of Asymmetric Conflict* 4.2, pp. 92–102. DOI: 10.1080/17467586.2011.627932. eprint: https://doi.org/10.1080/17467586.2011.627932. URL: https://doi.org/10.1080/17467586.2011.627932.

Pennebaker, James W., Matthias R. Mehl and Kate G. Niederhoffer (2003). 'Psychological Aspects of Natural Language Use: Our Words, Our Selves'. In: *Annual Review of Psychology* 54.1. PMID: 12185209, pp. 547–577. DOI: 10.1146/annurev.psych.54.101601.145041. eprint: https://doi.org/10.1146/annurev.psych.54.101601.145041. URL: https://doi.org/10.1146/annurev.psych.54.101601.145041.

Reimers, Nils and Iryna Gurevych (2019). 'Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks'. In: *CoRR* abs/1908.10084. arXiv: 1908.10084. URL: http://arxiv.org/abs/1908.10084.

Rivera-Soto, Rafael A., Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop and Nicholas Andrews (2021). *Learning Universal Authorship Representations*. DOI: 10.18653/v1/2021.emnlp-main.70.

Roser, Max and Esteban Ortiz-Ospina (2016). 'Literacy'. In: *Our World in Data*. URL: https://ourworldindata.org/literacy.

Ross-Larson, Bruce (1999). *Powerful Paragraphs*. W.W. Norton & Company.

Ruder, Sebastian, Parsa Ghaffari and John G. Breslin (2016). 'Character-level and Multi-channel Convolutional Neural Networks for Large-scale Authorship Attribution'. In: *CoRR* abs/1609.06686. arXiv: 1609.06686. URL: http://arxiv.org/abs/1609.06686.

Rudman, Joseph (1997). 'The State of Authorship Attribution Studies: Some Problems and Solutions'. In: *Computers and the Humanities* 31.4, pp. 351–365. ISSN: 00104817. URL: http://www.jstor.org/stable/30200436 (visited on 26/06/2023).

Rybicki, Jan and Maciej Eder (2011). 'Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words?' In: *Literary and Linguistic Computing* 26.3, pp. 315–321. ISSN: 0268-1145. DOI: 10.1093/llc/fqr031. eprint: https://academic.oup.com/dsh/article-pdf/26/3/315/3955977/fqr031.pdf. URL: https://doi.org/10.1093/llc/fqr031.

Sebranek, Patrick, Dave Kemper and Verne Meyer (2006). *Writers Inc: A Student Handbook for Writing and Learning*. Write Source, Great Source Education Group.

Shannon, C. E. (1951). 'Prediction and Entropy of Printed English'. In: *The Bell System Technical Journal* 30.1, pp. 50–64. DOI: 10.1002/j.1538-7305.1951.tb01366.x.

Shrestha, Prasha, Sebastian Sierra, Fabio González, Manuel Montes, Paolo Rosso and Thamar Solorio (2017). 'Convolutional Neural Networks for Authorship Attribution of Short Texts'. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, pp. 669–674. URL: https://aclanthology.org/E17-2106.

Skovholt, Karianne, Anette Grønning and Anne Kankaanranta (2014). 'The Communicative Functions of Emoticons in Workplace E-Mails: :-)'. In: *Journal of Computer-Mediated Communication* 19.4, pp. 780–797. DOI: https://doi.org/10.1111/jcc4.12063. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcc4.12063. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jcc4.12063.

Stamatatos, Efstathios (2009). 'A Survey of Modern Authorship Attribution Methods'. In: *Journal of the American Society for Information Science and Technology* 60.3, 538–556. DOI: 10.1002/asi.21001.

— (2016). 'Authorship Verification: a Review of Recent Advances'. In: *Research in Computing Science* 123, pp. 9–25. DOI: 10.13053/rcs-123-1-1.

— (2017). 'Authorship Attribution using Text Distortion'. In: DOI: 10.18653/v1/e17-1107.

— (2018). 'Masking Topic-Related Information to Enhance Authorship Attribution'. In: *Journal of the Association for Information Science and Technology* 69.3, pp. 461–473. DOI: https://doi.org/10.1002/asi.23968. eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23968. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23968.

Strunk, William and E. B. White (1972). *The Elements of Style*. Macmillan.

Theophilo, Antonio, Romain Giot and Anderson Rocha (2021). 'Authorship Attribution of Social Media Messages'. In: *IEEE Transactions on Computational Social Systems*, pp. 1–14. DOI: 10.1109/TCSS.2021.3123895.

Thompson, Dominic and Ruth Filik (2016). 'Sarcasm in Written Communication: Emoticons are Efficient Markers of Intention'. In: *Journal of Computer-Mediated Communication* 21.2, pp. 105–120. ISSN: 1083-6101. DOI: 10.1111/jcc4.12156. eprint: https://academic.oup.com/jcmc/article-pdf/21/2/105/19491845/jjcmcom0105.pdf. URL: https://doi.org/10.1111/jcc4.12156.

Van den Besselaar, Peter and Charlie Mom (2022). 'The Effect of Writing Style on Success in Grant Applications'. In: *Journal of Informetrics* 16.1, p. 101257. ISSN: 1751-1577. DOI: https://doi.org/10.1016/j.joi.2022.101257. URL: https://www.sciencedirect.com/science/article/pii/S1751157722000098.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser and Illia Polosukhin (2017). 'Attention is All you Need'. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vig, Jesse (2019). 'A Multiscale Visualization of Attention in the Transformer Model'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Florence, Italy: Association for Computational Linguistics, pp. 37–42. DOI: 10.18653/v1/P19-3007. URL: https://aclanthology.org/P19-3007.

Vig, Jesse and Yonatan Belinkov (2019). 'Analyzing the Structure of Attention in a Transformer Language Model'. In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, pp. 63–76. DOI: 10.18653/v1/W19-4808. URL: https://aclanthology.org/W19-4808.

Wang, Chenguang, Mu Li and Alexander J. Smola (2019). 'Language Models with Transformers'. In: *CoRR* abs/1904.09408. arXiv: 1904.09408. URL: http://arxiv.org/abs/1904.09408.

Wang, Lingxiao, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang and Quanquan Gu (2020). 'Improving Neural Language Generation with Spectrum Control'. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=ByxY8CNtvr.

Wardhaugh, Ronald and Janet M Fuller (2021). *An Introduction to Sociolinguistics*. John Wiley & Sons.

Warren, Robert Penn (1952). *The Sound and the Fury: The Corrected Text*.

Wegmann, Anna and Dong Nguyen (2021). 'Does It Capture STEL? A Modular, Similarity-based Linguistic Style Evaluation Framework'. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 7109–7130. DOI: 10.18653/v1/2021.emnlp-main.569. URL: https://aclanthology.org/2021.emnlp-main.569.

Wegmann, Anna, Marijn Schraagen and Dong Nguyen (2022). 'Same Author or Just Same Topic? Towards Content-Independent Style Representations'. In: *Proceedings of the 7th Workshop on Representation Learning for NLP*. Dublin, Ireland: Association for Computational Linguistics, pp. 249–268. DOI: 10.18653/v1/2022.repl4nlp-1.26. URL: https://aclanthology.org/2022.repl4nlp-1.26.

Wikipedia (2023). *Blues*. URL: https://en.wikipedia.org/wiki/Blues.

Wolfradt, Uwe and Jean E. Pretz (2001). 'Individual Differences in Creativity: Personality, Story Writing, and Hobbies'. In: *European Journal of Personality* 15.4, pp. 297–310. DOI: 10.1002/per.409. eprint: https://doi.org/10.1002/per.409. URL: https://doi.org/10.1002/per.409.

Xiao, Xinyu, Lingfeng Wang, Kun Ding, Shiming Xiang and Chunhong Pan (2019). 'Deep Hierarchical Encoder–Decoder Network for Image Captioning'. In: *IEEE Transactions on Multimedia* 21.11, pp. 2942–2956. DOI: 10.1109/TMM.2019.2915033.

Xie, Yaochen (2008). 'Hemingway's Language Style and Writing Techniques in" The Old Man and the Sea".' In: *English Language Teaching* 1.2, pp. 156–158.

Yang, Shuiqing, Chuanmei Zhou and Yuangao Chen (2021). 'Do Topic Consistency and Linguistic Style Similarity Affect Online Review Helpfulness? An Elaboration Likelihood Model Perspective'. In: *Information Processing & Management* 58.3, p. 102521. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2021.102521. URL: https://www.sciencedirect.com/science/article/pii/S0306457321000303.

Zhao, Chen, Wei Song, Xianjun Liu, Lizhen Liu and Xinlei Zhao (2018). 'Research on Authorship Attribution of Article Fragments via RNNs'. In: *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, pp. 156–159. DOI: 10.1109/ICSESS.2018.8663814.

Zhao, Ying and Justin Zobel (2007). 'Searching with style: Authorship Attribution in Classic Literature'. In: *Proceedings of the thirtieth Australasian conference on Computer science-Volume 62*. Citeseer, pp. 59–68.

Zhu, Jian and David Jurgens (2021). *Idiosyncratic but not Arbitrary: Learning Idiolects in Online Registers Reveals Distinctive yet Consistent Individual Styles*. DOI: 10.48550/ARXIV.2109.03158. URL: https://arxiv.org/abs/2109.03158.