Uncovering the Authors: A Dual Approach to Detect and Understand Authorship in Text Documents

Authors

Stephan Nef
Lukas Bamert

Supervisor's
Prof. Dr. Siegfried Handschuh
M.Sc Bernhard Bermeitinger

8,862 Natural Language Processing with Deep Learning
University of St. Gallen

https://github.com/bamertl/nlp_deep_project
https://universitystgallen.streamlit.app

Contents

1 Introduction

In academia, the integrity and originality of student work is of paramount importance. Ensuring that written assignments and research papers are authored by students themselves is essential to maintaining academic integrity. A burgeoning area of research, authorship attribution, has shown its usefulness in analysing textual documents to distinguish writing styles and author identities, at least in the current mode of teaching, but is likely to be heavily adapted in the future. This project focuses on academic papers written by students and aims to develop a sophisticated dual approach to authorship detection and analysis. The first phase involves the use of a deep learning model trained to detect changes in authorship within a document. The second phase uses linguistic feature engineering to provide interpretability by analysing the linguistic features that characterise an author's style. Finally, both approaches will be integrated. In order to effectively evaluate the proposed dual approach to authorship attribution, it was imperative to use a dataset that accurately represents the type of academic documents the system is designed to analyse. While there are publicly available datasets [1], these may not be representative of the student-authored documents in the context in which the system will be used. However, the public dataset was of great importance for training the deep model. For subsequent evaluation, a custom dataset was created as part of this project. This dataset, consisting of academic texts written by students, was designed to closely mimic the real-world scenarios in which the system would be used. This allowed for a more relevant evaluation of the developed models and a deeper understanding of the nuances involved in authorship attribution in the specific context of academic environments. The system was developed in Python and made available through Streamlit [2].

1.1 History of author attribution

The first attempts to distinguish authors on the basis of written text began in the 19th century with Mendenhall [3]. Mendenhall's work is considered seminal in the field, focusing on the quantitative analysis of word lengths in Shakespeare's plays [4]. In the 20th century, statistical methods began to be used [5]. By the turn of the 21st century, up to a thousand stylistic features were said to have been measured, examples being sentence length, word length and various frequencies [3]. With the advent of transformers, neural networks such as Bert and its descendants are now used to identify authors [6]. Deep models give the best results, but it is very difficult to interpret and reproduce these results. In generative models, chain-of-thought has been used to improve the interpretability of the models, which, in addition to interpretation, also improves the performance of the model. However, this does not provide a true interpretation of the system-internal decisions [7]. In addition, techniques such as Chain-Of-Thought are not currently available for non-generative models.

2 Method

2.1 Deep Learning Model Training

In the first phase, the focus is on using deep learning models to autonomously detect changes in authorship within scientific documents. Transformers have revolutionised linguistic task solving [8]. There are now language models that, on average, have a better understanding of language than humans [9]. Generative Transformer models, which have only the decoder part of the traditional Transformer, are particularly

popular. An example of this is ChatGPT [10]. However, generative models are trained to predict the next token based only on previous input, which is not appropriate in our case. For this reason, we used slightly older Bert-like models [11]. These models allow us to logically separate input sequences using Ids. After training several BERT-like models on the dataset of academic texts with known author transitions, we evaluated their performance in detecting authorship changes. In this project, we chose state-of-the-art models of BERT successors, namely RoBERTa and DistilBERT [12] [13]. In each case, pre-trained models that can be downloaded from Huggingface were used and trained with the Huggingface library [14]. The DistilBERT and RoBERTa models used were both pre-trained on Wikipedia data and the book corpus [15] [16].

2.1.1 Dataset

To train the model, we were given access to the PAN22 dataset [1]. The data in this dataset was collected and compiled from various sites on the Stackexchange network. The dataset consists of three subsets, each created for a different purpose. The first subset, Style Change Basic, is limited to text written by two authors and contains only one author change. The second, Style Change Advanced, consists of text that has been written by more than one author and therefore contains multiple author changes. The last, the Style Change Real-World, is the same as the second, except that author changes occur not only by paragraph, but also by sentence. The dataset had to be pre-processed in order to train a Bert model. As mentioned above, the model takes two sequences as input and outputs the probability of whether an author change has taken place. For the training of our model for our purpose, all data sets were therefore mixed and split into inputs of two sequences each.

2.2 Linguistic Feature Engineering

The second phase of the project is about understanding the linguistic features that characterise an author's writing style. This is a kind of fingerprint. This is crucial because deep learning models, while very powerful, are often described as "black boxes": They are not interpretable. The development of linguistic features is a careful process. It focuses on the extraction and systematic analysis of linguistic features.

First, syntactic features are examined. Syntax refers to the arrangement of words and phrases into well-formed sentences. Different authors may have different preferences for structuring their sentences. For example, some authors prefer complex sentences with multiple clauses, while others use simpler sentence structures. Common metrics such as the Flesch-Kincaid score, the Gunning-Fog index or the Coleman-Liau index are used to assess the readability of a text.
Indications of the author's syntactic preferences are obtained by extracting features such as the average sentence length, the proportion of different clause types (independent, dependent) and the complexity of sentence structures. Lexical features are also analysed. These refer to an author's choice of words and vocabulary. Lexical variety, i.e. the variety of words used by an author, is an important indicator of style. The use of so-called n-grams is one aspect of understanding an author's writing style. N-grams are successive sequences of n words in a text. By analysing the frequency and distribution of n-grams, we can gain insights into an author's linguistic patterns and preferences. For example, certain phrases or word combinations are often used by certain authors and contribute to their characteristic style. By studying n-grams, we can identify these recurring patterns and further improve our understanding of the author's writing style. The inclusion of N-grams in the process of developing linguistic features allows for a more comprehensive analysis of an author's style.

The frequency of certain words or phrases, as well as the use of specialised terminology, can also provide clues to authorship. Grammatical structures are also assessed. This includes the use of word types such as nouns, verbs, adjectives and their arrangement in sentences, and can also provide useful insights in combination with semantic analysis. The analysis of grammatical patterns, such as the frequency of passive sentences compared to active sentences or the occurrence of certain grammatical constructions, can help to characterise an author's writing style. Another important element is the analysis of punctuation and formatting. Different authors may show different patterns in the use of punctuation marks such as commas, semicolons and parentheses. In addition, formatting preferences such as paragraph length and indentation style can provide clues to authorship.

Using a comprehensive range of linguistic features, from syntactic and lexical aspects to the study of n-grams, grammatical structures and punctuation/formatting patterns, a multidimensional understanding of an author's writing style emerges. This methodological approach paves the way for improved interpretation and deeper insights into the intricate fabric of an author's unique literary voice.

3 Results

3.1 Part 1: Deep Learning Model Performance

The pre-trained models of the type RoBERTa and DistilBERT were used. Due to the large training time, the training had to be split over different days, as Google Colab has a maximum time of free training [17]. For this reason, unfortunately, no hyperparameter search could be carried out. Especially the RoBERTa training turned out to be extremely slow, which is why we quickly concentrated mainly on DistilBERT. The best model, based on DistilBERT, achieved an accuracy of 69.94% on the validation data. This is surprising, as for us personally, the author changes were almost undetectable. The following Figure 1 visualises the confusion matrix of the best model evaluated on the validation data of the processed PAN22 dataset.

In the following Figure 2 we display the metrics of the training of the best performing model.

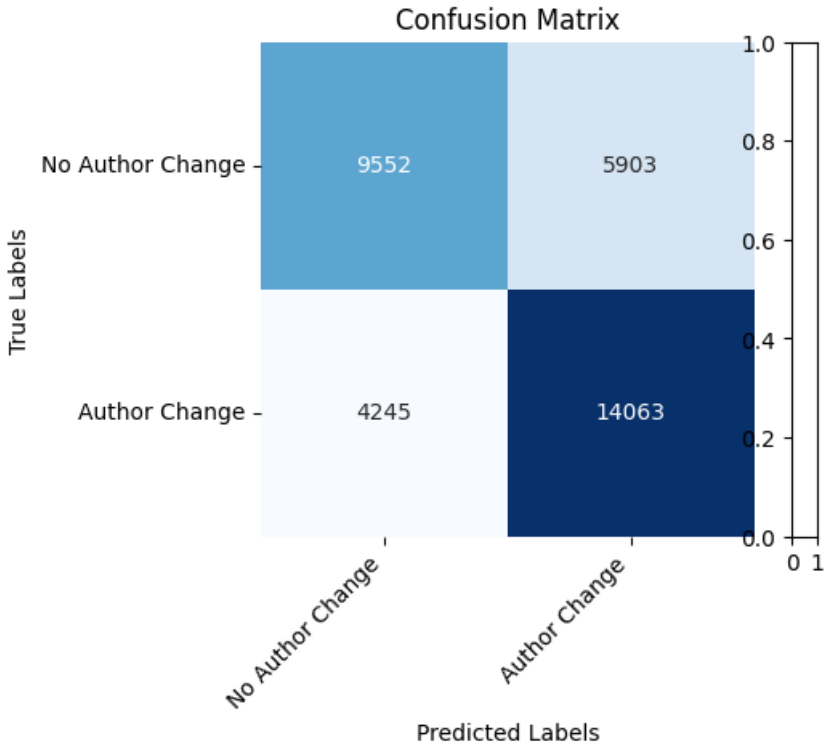

Figure 1: Confusion Matrix of the best model

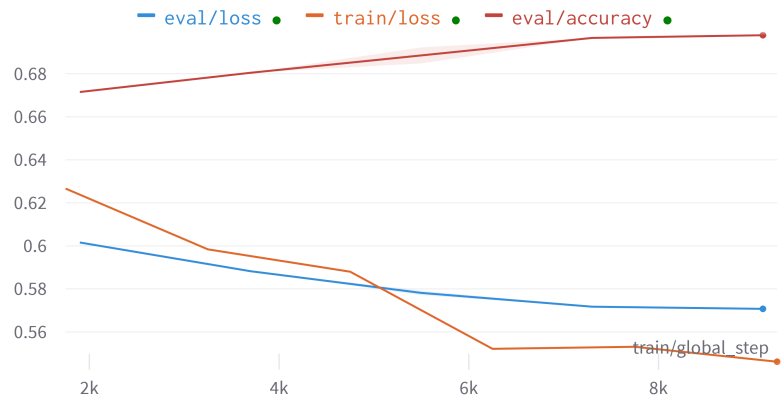

Figure 2: Metric of best model

In the following Figure 3, there is an example displayed of the model used on a students work.

This can be for papers on the application directly. Most of the time it actually does not predict an author change in a document of the same author.

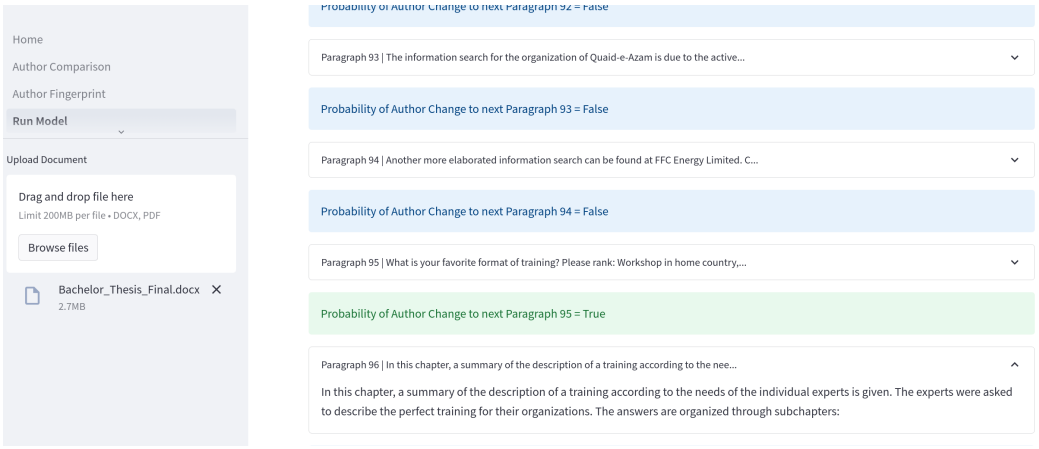

Figure 3: Author Change on Student Work

3.2 Part 2: Linguistic Feature Insights

The linguistic feature engineering phase resulted in a rich set of attributes characterizing authorial styles. For instance, distinctive patterns in sentence structure and punctuation use were observed among different authors. Through statistical analysis, certain features were identified as highly indicative of authorship. These insights not only contribute to author identification but also provide a tangible understanding of the stylistic elements that constitute an author's unique writing style.

3.2.1 Imbalanced use of punctuation in ChatGPT texts compared to human-written texts

Texts generated by ChatGPT show a notable imbalance in the use of punctuation, especially commas. While human-written documents show a balanced distribution of about 60% +- 5% (in 95% confidence interval) for comma usage, ChatGPT-generated text shows a higher ratio of 75%. In Figure 4 the comparison across authors clearly shows the imbalance, and in Figure 5 the punctuation behaviour of the author Olivia Frigo-Charles shows a certain stability over all analysed documents.

This discrepancy can be attributed to several factors. ChatGPT's training data, derived from web text, often reflects informal and conversational styles that tend to overuse commas for pauses and flow. In addition, the model's reliance on statistical patterns learned from the training data tends to replicate the observed overuse of commas. Furthermore, the lack of contextual understanding in ChatGPT results in the default use of commas without accurately assessing sentence structure or meaning.
The imbalance in punctuation affects readability and clarity, as run-on sentences and convoluted phrases can confuse readers and disrupt the flow of ideas. Addressing this issue requires refining training with more balanced data and incorporating contextual understanding to improve punctuation choices based on the intended meaning.
To conclude, the higher punctuation imbalance, especially the overuse of commas in ChatGPT-generated text, stems from informal training data, statistical patterns and limited contextual understanding. Careful review and editing is required to ensure optimal readability and coherence in the final output.

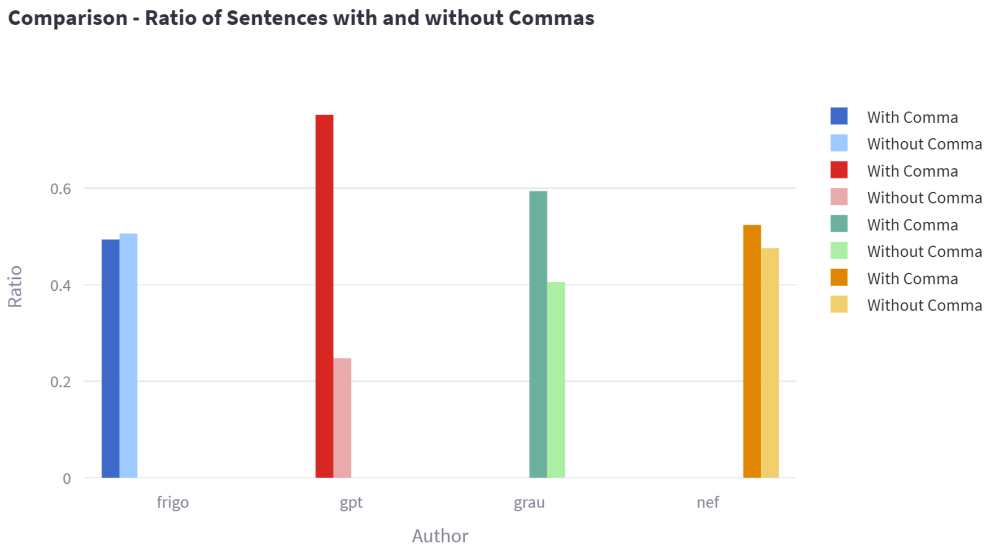

Figure 4: Analysis of punctuation style per author

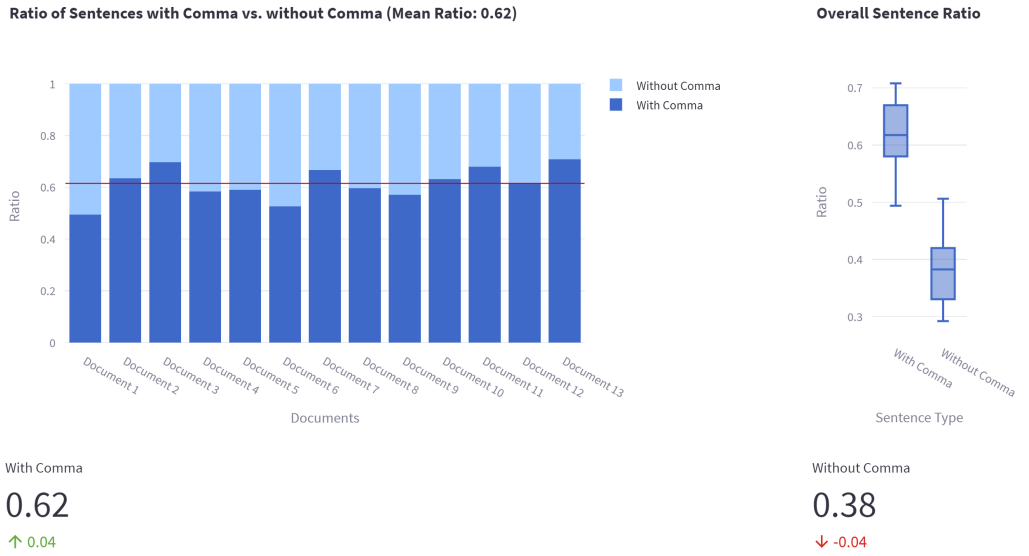

Figure 5: Analysis of punctuation behaviour of the author Olivia Frigo-Charles over 12 different own written documents without any help of ChatGPT.

3.2.2 Analysis of Flesch-Kincaid Grade Levels and Type-Token Ratio

The Flesch-Kincaid Grade Level is a metric used to determine the readability of a text based on its sentence structure and word complexity. Analysing the Flesch-Kincaid Grade Level results for the different authors, we observe the levels for author Frigo: 9.5, ChatGPT: 14.9, Grau: 11.1 and Nef: 9.4 as shown in Figure 6.
The Flesch-Kincaid Grade Level represents the number of years of education required to understand a piece of text. A lower grade level indicates easier readability, while a higher grade level indicates more complex and advanced writing. Comparing the results, we find that ChatGPT has the highest Flesch-Kincaid Grade Level, indicating that the text is more difficult to understand. There are a number of possible reasons

why ChatGPT's grade level might be higher:
ChatGPT's speech generation is based on a large dataset with a diverse vocabulary, including both complex and less common words. As a result, the use of such vocabulary contributes to a higher level of proficiency. In addition, the model shows a variety of sentence structures, often with longer and more complicated constructions. This presence of complex sentence patterns further increases the level of proficiency, making the text more challenging to read and understand. Despite being exposed to large amounts of training data, ChatGPT may not fully understand the context or intended meaning behind the generated content. As a result, it may produce syntactically correct sentences that require a higher level of comprehension to interpret effectively.
It's important to note that the Flesch-Kincaid Grade Level is not necessarily an indicator of writing quality. While ChatGPT may have a higher grade level, it doesn't necessarily mean it has better content or information. Grade level should be considered in the context of the intended audience and purpose of the text. Additionally, it is interesting to explore other metrics, such as the Type Token Ratio, to gain further insights into the text's lexical diversity and potential impact on readability. As can be seen in Figure 7, there is nothing out of the ordinary in the comparison of authors, but for an author's fingerprint, it is still a valuable feature to consider the author's writing style, creativity and target audience.

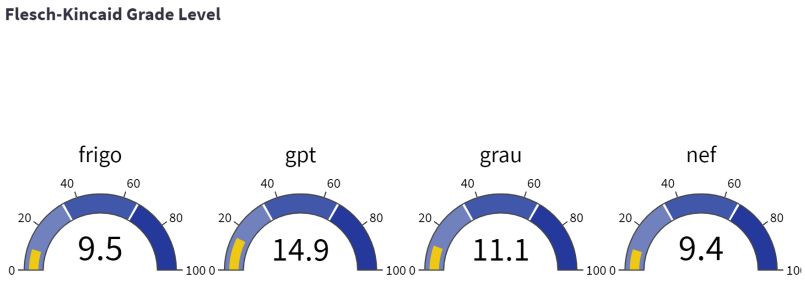

Figure 6: The Flesch-Kincaid Grade Level is a readability test that indicates the approximate grade level required to understand a text. A lower grade level indicates easier readability, while a higher grade level suggests more advanced vocabulary and sentence complexity.

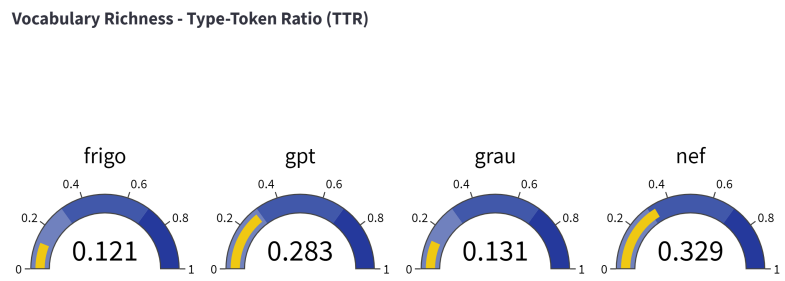

Figure 7: The Type-Token Ratio (TTR) is a measure of vocabulary richness that calculates the ratio of unique words (types) to the total number of words (tokens) in a text. A higher TTR indicates a larger variety of words used, suggesting a more diverse and potentially sophisticated vocabulary.

4 Discussion and Outlook

An important challenge faced during this project was the collection of actual student-authored works containing author changes. Gathering real-world academic documents with known author changes is intrinsically difficult due to several reasons. This includes privacy concerns and the general scarcity of such documents. Furthermore, it is essential to consider that student works are often proofread before submission. This proofreading process, whether carried out by peers, tutors, or automated tools, can modify the original writing style of the student, making authorship detection more challenging. Proofreading can introduce stylistic alterations that might be incorrectly attributed to author changes by the system. In addition, technological advancements and tools have implications for authorship attribution. Students have been utilizing tools like Deepl for translations, which can alter the inherent style of the text. Moreover, with the advent of sophisticated language models like ChatGPT, it is likely that an increasing number of students might employ these tools for generating or enhancing their academic texts. These tools could significantly transform the writing style, and their widespread adoption may render the creation of reliable datasets with author changes nearly impossible in the future. Another facet to consider is the multilingual aspect of student writing. Students who are proficient in multiple languages might compose academic documents in different languages. The writing style of an individual can vary significantly between languages due to differences in grammar, syntax, and cultural nuances. This variation adds an additional layer of complexity to the task of authorship attribution and necessitates the consideration of language-specific features and models. Considering these challenges and intricacies, it is imperative for future research and development in authorship attribution to be adaptive and cognizant of the evolving landscape of academic writing. The incorporation of robust methods for handling proofread texts, translations, and multilingual authorship, along with a critical evaluation of the impact of advanced language models on authorship, will be essential for the continued effectiveness and relevance of authorship attribution systems in academia.

References

[1] E. Zangerle, M. Mayerl, M. Tschuggnall, M. Potthast, and B. Stein, "PAN22 Authorship Analysis: Style Change Detection." Zenodo, Mar. 2022.

[2] "Streamlit." `https://streamlit.io/`. Accessed on 2023-06-18.

[3] E. Stamatatos, "A survey of modern authorship attribution methods," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, pp. 538–556, 2009.

[4] T. C. Mendenhall, "The characteristic curves of composition," *Science (New York, N.Y.)*, vol. 9, pp. 237–246, 1887.

[5] G. U. Yule, "On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship," *Biometrika*, vol. 30, no. 3/4, pp. 363–390, 1939.

[6] A. Abbasi, A. R. Javed, F. Iqbal, Z. Jalil, T. R. Gadekallu, and N. Kryvinska, "Authorship identification using ensemble learning," *Scientific reports*, vol. 12, no. 1, p. 9537, 2022.

[7] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[9] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, and B. Ge, "Summary of chatgpt/gpt-4 research and perspective towards the future of large language models," 2023.

[10] "Chatgpt." `https://chat.openai.com/`. Accessed on 2022-06-10.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[13] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2020.

[14] "Hugging face – the ai community building the future.." `https://huggingface.co/`. Accessed on 2023-06-18.

[15] W. Foundation, "Wikimedia downloads." `https://dumps.wikimedia.org`. Accessed on 2023-06-18.

[16] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[17] "Google colab." `https://colab.research.google.com/signup`. Accessed on 2023-06-18.

Directory of aids

I hereby declare that I have used Deepl, Grammarly and ChatGPT as an aid for translation, grammar correction and proofreading.

Declaration of authorship

Lukas Bamert, Stephan Nef

"We hereby declare

- that we have written this writing sample (thesis, seminar paper, term paper or other) without any help from others and without the use of documents and aids other than those stated in the references or directory of aids,
- that we have mentioned all the sources used and that I have cited them correctly according to established academic citation rules,
- that the topic or parts of it are not already the object of any work or examination of another course unless this is explicitly stated,
- that we am aware that my work can be electronically checked for plagiarism and that we hereby grant the University of St.Gallen copyright as far as this is required for this admin-istrative action."