Identifying Crosswriters' Altering Style in Books for Children and Adults Using Supervised Machine Learning

Dimitris Bouparis
dimitrios.bouparis@student.uantwerpen.be

Abstract

Stylometry is the quantitative study of literary style through computational distant reading methods. It is based on the observation that authors tend to write in relatively consistent, recognisable, and unique ways (Laramée, 2018). Identifying the similarities and differences in style, content, and genre between literature intended for children and adults has always been under the radar of researchers in the field of Computational Literary Studies. However, only recently has examining the implications of cross-writing (i.e., writing works for various readership age groups) gotten attention. In this study, supervised machine learning methods were applied to get a better understanding on whether and how such authors ("crosswriters") alter their style when targeting a different age group, based entirely on content words. The study was conducted on 5 English authors, and the SVM models reach an F1 macro score of .74 when predicting the age group using all texts and .93 on average for each of the authors individually.[1] To achieve these results, it was essential to overcome the issue of overfitting on the characters of the stories, which was made possible by implementing a Named Entity Recognition (NER) step in the preprocessing pipeline and leaving at least one book by each author out of the train set entirely in each of the 10 folds during Cross-Validation.

Keywords: *computational stylometry*, *crosswriters*, *children's literature*, *digital humanities, support vector machines*.

1. Introduction

Recent advancements in computing, combined with the statistical revolution and the enormous amount of daily data generation thanks to accessing to technology getting easier by the day, changed the way science works and research is conducted. The Humanities did not escape the trend, albeit being slow to adapt and adopt it, and started implementing digital tools, giving a whole new aura to making use of data to revisit old studies and, of course, pave new ways for topics we never thought possible (Berry,

[1] All the code is available on https://github.com/dimboump/crosswriters/. The texts and the metadata are under copyright and, as a result, not available.

2012; Gold, 2012). This ‘quantitative turn’, however, had already started much earlier, prior to falling out of favour in the 1980s following the ‘cultural turn’ (Karsdorp et al., 2021). One of the most interesting fields that machine learning and statistics have given new life to is Literary Studies. Traditionally, research in this field was relying on large groups of people annotating a few texts or, worse, few people annotating large amounts of text. More recently, conducting studies in works written by crosswriters has gained interest. According to Beckett (1999, xi-xii), the term ‘crosswriters’ can be traced back to 1993 and refers to authors who write separate works for readers of various ages. The recent spark in interest in such studies can be attributed to two main reasons: (a) the ability to remove the layer of additionally comparing different authors’ idiolects and instead focusing on the style of the same author targeting different age groups; and (b) the increasing availability and use of computational means and methods to do the heavy lifting of finding underlying patterns where humans find meaning and personal opinions. In the latter case, the field of (Computational) Stylometry is the one that can shed light on whether and, if so, how the style of a crosswriter differs among the intended readerships. As McIntyre and Walker (2019: 65) have pointed out, relying on computational stylometry shifts the scholars’ primary focus on certain stylistic features to more ‘background features of authorial style’.

It is essential to point out that stylometry – regardless of computationally or traditionally approached – is based on the notion of the so-called ‘stylome hypothesis’. According to van Halteren et al. (2005), ‘authors can be distinguished by measuring specific properties of their writings, their stylome as it were’. A clear association between the stylome and the genome cannot be overlooked, as it is believed that both are unique.

The aim of this paper is to give a satisfactory answer to the questions: Do crosswriters alter their writing style depending on the age of their intended readers? If so, do content words influence this at a significant enough level to only focus thereon? To approach these questions, we first need to justify the use of computational means. Computer-aided Stylometry, as it is called, can (a) eliminate the subjective component that is inherent to close reading; and (b) give a better understanding of the deeper choices that an author, consciously or not, has made when producing works targeting different age groups. A supervised machine learning method was used to tackle this problem.

2. Related work

It is no secret that the writing style of every one of us is the accumulation of our lexical and syntactical choices. What has yet to be (dis)proven is whether these choices are made unconsciously or not, and, if so, what the main features are that do not 'survive' the transition between addressing different ages of readers. This study falls under the umbrella of genre identification, as it tries to classify texts as either meant to be read by children or (young) adults. To this end, there are two main paths other researchers have followed until now: focusing either on grammatical and syntactical aspects of the texts in question or solely on lexical statistics.

2.1 Grammatical and syntactical features

Focusing on grammar and syntax used to be the dominant paradigm in stylometry. For example, counting the frequency of passive voice use and part-of-speech tagging were among the most used features in identifying an author's style – and then using these data to distinguish one author from another. The main drawback of such approaches, and mainly of syntactic features, is that the corpus needs to be annotated. This limited the potential lengths of the corpora and at the same time increased not only the complexity but also the cost of such projects, while also introducing an element of subjectiveness, since annotators tend to disagree in some cases (Davani et al., 2022). Some approaches (Stamatatos et al., 2001; Finn & Kushmerick, 2006; van Halteren, 2007; Grieve, 2007) include taking into consideration both low-level measures (e.g., sentence length, punctuation mark count, etc.) and syntax-based ones (e.g., noun phrase count, verb phrase count etc.) to capture stylistic information. Others (Wu et al., 2021; Zheng & Jin, 2022) combined traditional statistical methods (average word/sentence length, frequency of digits, and others) with modern, multichannel self-attention mechanisms to assign importance weights to input features in order to select features and reduce noise. This leaves the options for model selection open and allows for mixing of already known features as well as newly discovered ones.

2.2 Lexical features

On the other hand, there is much prior work that did not focus on the grammatical or syntactical aspects of writing. Instead, they showed that the way the function words are used (Argamon & Shlomo, 2005; Zhao & Zobel, 2006; Kestemont, 2014) or that simple

character or word *n*-grams (Luyckx & Daelemans, 2011; Gomez Adorno et al., 2018) can also be indicative of the style an author writes in and, therefore, significantly helpful in authorship attribution tasks. Others suggest focusing on the number of words that occur only once or twice in the texts (Stamatatos, 2007), the type-token ratio (López-Escobedo et al., 2018) or the lexical richness (Mikros & Argyri, 2007). However, this method, as Sichel (1975) showed, can be very unreliable in texts of length less than 1,000 words. In order to isolate the effects of topic and genre, Luyckx and Daelemans (2011) collected 300 genre-controlled texts distributed in 3 author categories (2 individual authors and 1 "Others" category) consisted of texts of 10 different authors. In their paper, they also included some collaborative articles of the aforementioned two authors). The common denominator in all the above studies is the employment of at least one Bag-of-Words (BoW) approach. In other words, the absolute frequency of each word (or, rather, lemma) is added to a Document-Term Matrix (DTM). Interestingly enough, the hyperparameters of the models regarding the minimum and the maximum document frequency threshold are not provided in any of the papers. We therefore assume that all words (terms) were included in the DTM, regardless of whether they appeared once (*hapax legomena*), twice (*dislegomena*) or even in all texts (documents).

Finally, the most relevant prior study to the one at hand, as it focuses in particular on crosswriters, is the one conducted by Haverals et al. (2022), in which the authors examined 5 Dutch and 5 English crosswriters. Conversely to the present study, Haverals et al. used unsupervised stylometric techniques, mainly dimensionality reduction algorithms (Principal Component Analysis – PCA, and Hierarchical Cluster Analysis – HCA). In other words, the algorithms they implemented were given no extra information (e.g., the name of the author or the intended readership) apart from the very texts (Haverals et al., 2022: 18). In addition, they used the 100 most frequent words (MFW) as features and experimented with raising the sample to 300, although without getting significantly different results back. In their Conclusion, they raise the question, among others, whether focusing on content words would yield different, or even better, results. The study at hand explores this possibility, by removing the most common function words (*stopwords*) and using content words as features instead. We will elaborate on the methodology in Section 4. In short, the main findings of Haverals et al. can be summarised into three main points (2022: 19): (a) There is a high correlation

between the stylistic features of the author and the intended readers' age rather than the authors' age at the time of writing or the genre; (b) The books targeting young adults are stylistically more similar to the ones addressed to adults than to children; and (c) Some adult titles cluster with children's books for young readers due to the author employing a simpler style.

3. Corpus

The original corpus[2] consists of 753 works by 31 Dutch and English authors in total and was also used by Haverals et al. (2022). In the present study, however, only English works were paid attention to. In particular, the works by 5 authors (David Almond, Anne Fine, Neil Gaiman, Philip Pullman, and J.K. Rowling) were used, as they are the only ones who have published for both children and adult audiences. In addition, a few works co-authored by some of these crosswriters were not taken into consideration, as this would affect the stylometric signal. The final corpus has a total of 7,338,360 words. The number of available works per author for each age group can be seen in Figure 1.

3.1 Metadata information

A metadata file[3] was accompanying the corpus containing relevant information about the books and the authors, such as the year of publication, the gender of the author, the publisher's name but more importantly the intended readers' age, as found in the books' listings either in the publishers' or sellers' website (Haverals et al., 2022: 6). Haverals et al. divided the age groups as such: middle child (ages 6 to 8), late child (ages 9 to 11), young adult (ages 12 to 18), adult (18 and above). In our case, the division was initially done in line with them, however, a binary classification approach (child vs adult) was later preferred, as early results on the multiclass approach were not indicative of a clear separation among the younger age groups.

First, the metadata file was cleaned up to extract a baseline sample, containing only English single authors (conversly to collaborative works), and books the intended

[2] Many of the Dutch texts, although not used in the current study, were provided by the publishing house Querido and the authors themselves. The rest were scanned, digitised, and corrected by interns working on the Constructing Age for Young Readers (CAFYR) project (Haverals et al., 2021: 19-20). Learn more at https://cafyr.uantwerpen.be/en/.

[3] The metadata were also collected as part of the CAFYR project (see previous footnote).

readership thereof was mentioned in the respective column. Next, the final sample was derived by narrowing down further and including only authors who have published for both children and adults, ultimately converting the problem into binary. The number of total adult books therein is as little as 20% of the ones intended for children (Figure 1).

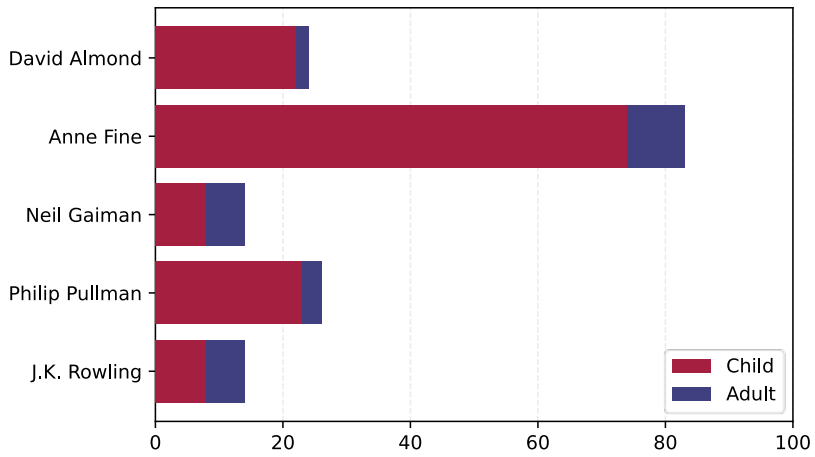

Figure 1: The number of books for each author per reader's age group in the corpus.

That introduces imbalance between the two target classes, but is to be expected, due to only Pullman and Gaiman writing for adults in their early careers, as we can conclude from the timeline-like visualisation of the number of books written by each author for each age group (Figure 2). An interesting observation is that all 5 authors have written more books intended for minors, with some even having 10 times as many books for minors compared to adults. An overview of the final corpus is shown in Table 1.

	Books			**Excerpts**		
	Children	**Adults**	**Total**	**Train**	**Test**	**Total**
David Almond	22	2	**24**	620	155	**775**
Anne Fine	74	9	**83**	1,237	310	**1,547**
Neil Gaiman	8	6	**14**	553	139	**692**
Phillip Pullman	23	3	**26**	1,280	320	**1,600**
J.K. Rowling	8	6	**14**	1,804	451	**2,255**
Total	**135**	**26**	**166**	**5,495**[4]	**1,374**	**6,869**

Table 1: Overview of the corpus as it was formed after deriving (a) the final sample, containing authors with books for both adults and children; and (b) after splitting into train and test subsets for the models.

[4] This number falls short by 1 if we add all the train excerpts. This is caused due to rounding taking place while train-test splitting in Scikit-Learn if we set the value of `test_size` to a float (see https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html).

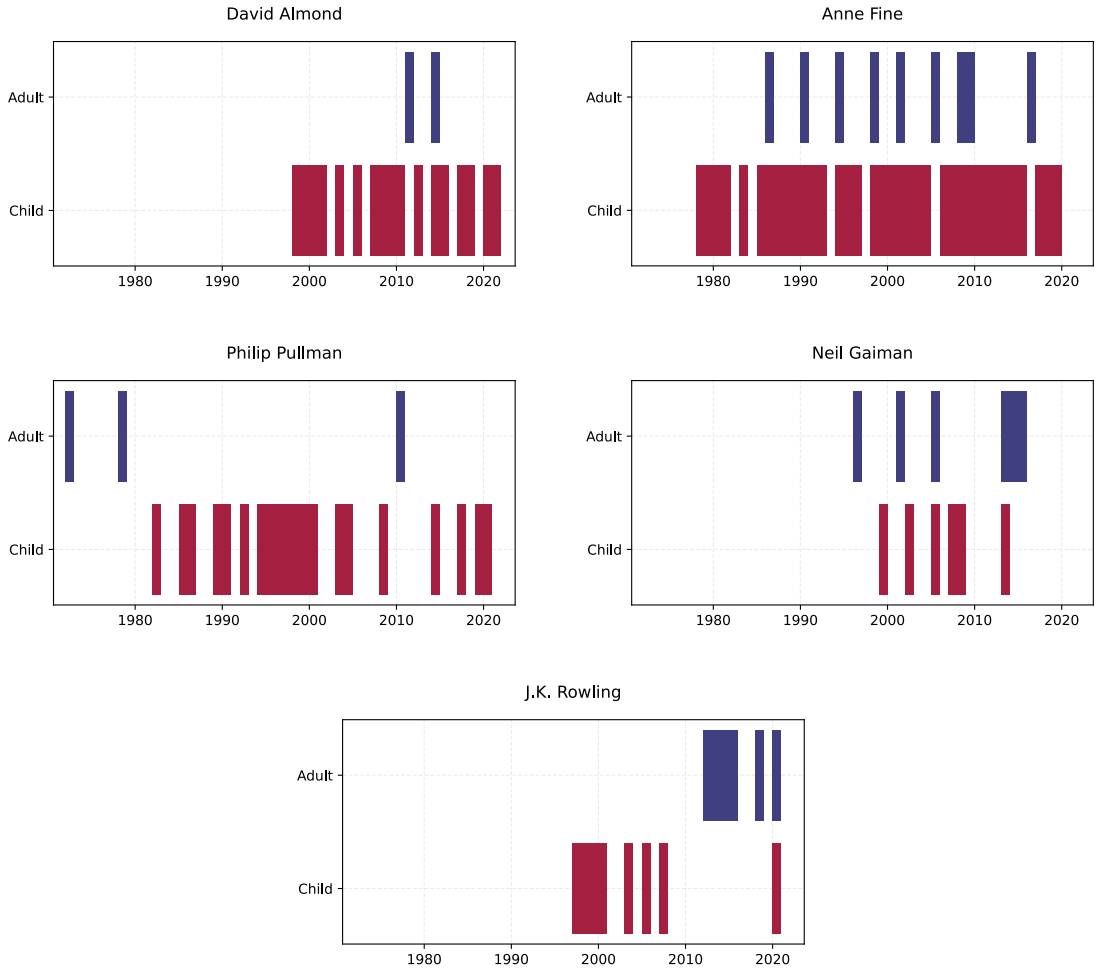

Figure 2: A timeline-like visualisation of the years the authors in question published at least one book per age group.

3.2 Preprocessing

After the final sample was extracted, it was time to deal with the actual texts. They initially underwent light preprocessing, namely lowercasing and tokenization. It is crucial to mention that they were also loaded in excerpts of 1,000 words. This number was not chosen arbitrarily, as it was suggested by plotting the total words distribution of the texts (Figure 3). Since no comparison among the authors themselves is going to be made – instead, how they themselves write for the two different age groups – it is not relevant whether authors are over or under-represented in the corpus.

With that in mind, the loaded corpus resulted in 6,869 segments from 166 files in total. The total number of segments for each author per age group is shown in Figure 4. Right from the start, we can clearly spot outliers on the right-hand side of the plot – which, almost certainly, are the *Harry Potter* series by J.K. Rowling. The plot is also skewed on the right, due to the outliers being present. The main reason the 1,000-word length was selected was to not exclude a large number of books in the process of loading the corpus. Then, the workflow was split to accommodate for the two different machine learning models.

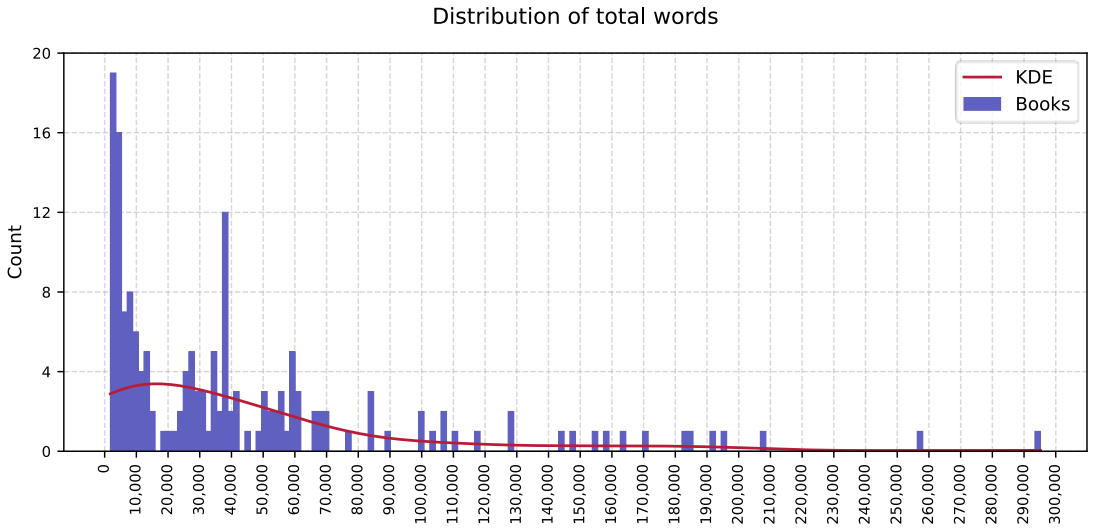

Figure 3: Distribution of the total words in the corpus. The Kernel Density Estimation (KDE) is also plotted (red line) to indicate what the underlying distribution of the words in the books might look like.

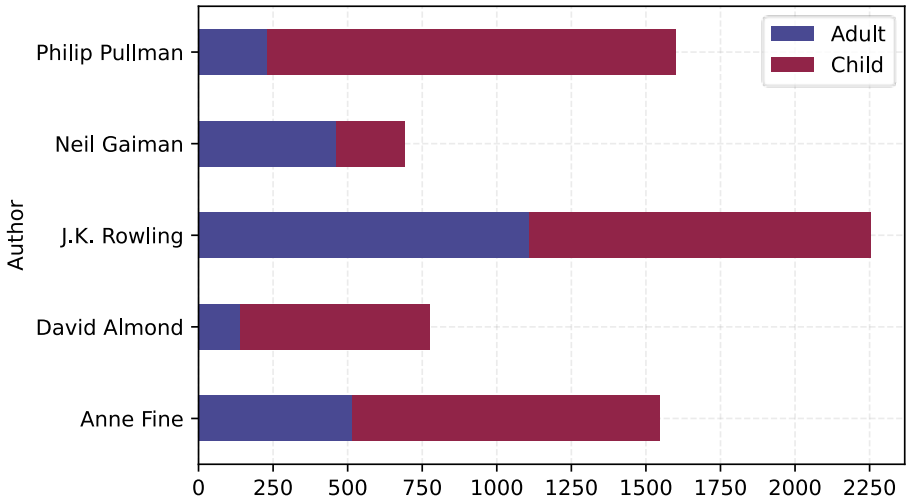

Figure 4: Total number of segments for each of the authors per intended readership age group.

4. Methodology

As pointed out earlier, the main objective of this paper is to examine whether it is possible to distinguish works written by the same author intended for minors and adults. By extension, we are solely focusing on what content words change as part of this transition. To achieve this, two separate Linear Support Vector Machines (SVM) models were implemented with 10-fold Grid Search Cross-Validation for parameter tuning. In both cases, the data were split into 80% training and 20% test, respectively, albeit following different methods (see subsections below).

4.1 All authors

To work with all English texts, we applied vectorization by employing Scikit-Learn's `CountVectorizer` and `TfidfVectorizer`[5] and using the same set of parameters for the SVM model. The Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer ultimately yields better results. However, as it will be highlighted in the next subsection, the models were significantly overfit. To avoid leakage of segments of the same work between the training and test data, a Leave-One-Text-Out approach was introduced by employing Scikit-Learn's `GroupShuffleSplit`[6]. In particular, at least one book by each author was excluded from the training subset and therefore only used as test data.

The SVM hyperparameters were determined after running Grid Search, with different values for the maximum document frequency set to .25, .50, and .75. Finally, the class weight parameter was tested with values set to `balanced` or double in favour of adults (.67 vs .33 respectively) to account for the imbalance of the two target classes. There was no need for the vectorizer to lowercase the input since the corpus was already lowercased during the preprocessing step. The kernel used is linear.

4.2 Individual authors

For the next 5 models, one for each of the authors we examine, the parameters that were passed in the Grid Search consisted of using unigrams, bigrams, and trigrams, `balanced` or `None` for class weights, no restriction on the total number of features, and

[5] See Scikit-Learn's extensive documentation and tutorials on the `TfidfVectorizer`: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html.

[6] See Scikit-Learn's extensive documentation and tutorials on the `GroupShuffleSplit`: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GroupShuffleSplit.html.

maximum document frequency of .25. The scoring was set to both accuracy and F1 macro, but refit was done with only the latter in mind. In each case, a confusion matrix was plotted along with the 50 features with the highest positive and negative coefficients for the Linear SVM. This gives us an overview of the words that play the most important role in determining whether a text was written for targeting children or adults. The frozenset of stopwords[7] and the entities was passed into the TF-IDF vectorizer and made sure that almost no names of characters and locations were picked up during training and emerged as significant features for the models (see Section 5.1).

5. Results

The two SVMs initially scored very high accuracy and relatively high F1 macro scores. This raised suspicion of potential overfitting, namely that the models learned the data very well during training and that the test data were not all that different (which is to be expected, given their origin). This turned out to be the case as soon as the coefficients of the top 25 features per class (child or adult) were plotted. As it turns out, the models, especially the single-author ones learned to distinguish the books based on the names of the characters in the stories. Although not all plotted features were names of characters, this does not allow us to arrive at the conclusion that the authors write differently depending on the intended reader's age group, as the focus shifts to the plot of each book/series rather than the writing style of its author.

5.1 Named Entity Recognition

To tackle this problem, a Named Entity Recognition step was added to the pipeline, prior to splitting the books into chunks of 1,000 words each, as we described in Section 3.2. The idea was that, even with the leakage provision, names that span across series (e.g., *Harry Potter*) would still 'leak' from training to test subsets. With that in mind, a pre-trained language model (PLM) was used to extract the names of characters (e.g., "Liz", "Hermione", "Dracula", etc.), locations ("Ireland", "Thames", "Hogwarts", etc.), and organizations (companies, universities, churches, etc.). The model that was

[7] Found in Scikit-Learn's source code on GitHub (latest revision: https://github.com/scikit-learn/scikit-learn/blob/5fd66bc55f03740f395971abf1189b11594252b1/sklearn/feature_extraction/text.py#L194). The included words have been taken from the Glasgow Information Retrieval Group, and the original list can be found at http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words.

used is the BERT-large-NER[8], using HuggingFace's Transformers library[9]. All these unique entities were saved into a Python set that was then added to the Scikit-Learn's frozenset of stopwords[7] and was fed to the vectorizer in the two pipelines (all-author and per-author models). One small caveat is that these entities were tokenized into single words, as the Grid Search included the parameter for unigrams.

This extra step, along with the leakage provision, turned out to be vital for the overall performance and prevention of overfitting. High scores were still achieved, but the extreme coefficients of the named entities were almost gone. These entities, quite expectedly, had previously taken the place of the top positive and negative coefficients in the plots. Some entities, mostly fictional names or names coming from languages other than English (e.g., Dumbledore, Ursula, etc.), were not identified by the BERT model, even when using its large version. This required that they be added to the stopwords list manually prior to the final run of the models. A direct comparison of the models' scores before and after the implementation of the NER is shown in Table 2.

	Pre-NER		**Post-NER**	
	Acc.	**F1**	**Acc.**	**F1**
David Almond	.914	.795	**.933**	**.857**
Anne Fine	.843	.800	**.979**	**.976**
Neil Gaiman	**.939**	**.928**	.931	.916
Phillip Pullman	.918	.771	**.962**	**.906**
J.K. Rowling	.991	.991	**.999**	**.999**
All authors	**.994**	**.851**	.795[10]	.743[10]

Table 2: Accuracy and F1 macro scores for the models, both for all authors combined and for each author individually, before and after implementing the Named Entity Recognition step with BERT (highest scores per row are marked with bold).

Although the scores went even higher for all except for Gaiman and the all-authors-at-once model, the feature importance coefficients plots suggest that the classification was done entirely on lexical frequencies (see Figure 5 and the Appendix).

[8] Available on HuggingFace: https://huggingface.co/dslim/bert-large-NER.

[9] See the HuggingFace documentation on Transformers: https://huggingface.co/docs/transformers/index.

[10] The all-authors post-NER scores include the group shuffle splitting, hence the models are less prone to overfitting, which justifies the lower scores.

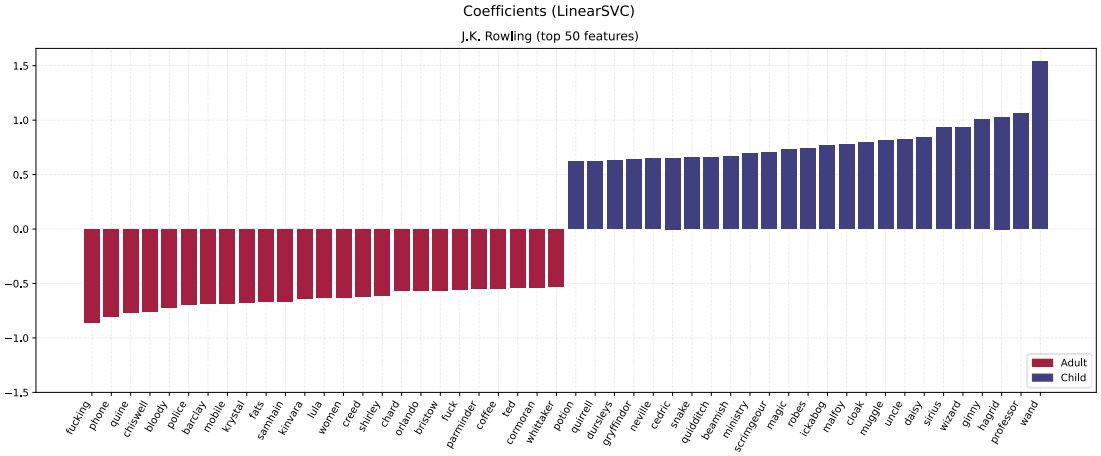

Figure 5: Plot of the feature coefficients of the Linear SVM model performed on J.K. Rowling's texts. Children's books were selected as the positive class (blue), therefore the negative coefficients (red) indicate the 25 most contributing features for identifying books targeting adults. The rest of the plots can be found in the Appendix.

5.2 Distinctive features

In all cases of the authors, the models preferred word unigrams as the optimal analyser. Therefore, single words were the determining features for the classification. In the top 50 features (i.e., 25 per age group), no lemma between the two categories is shared (see the Appendix). That is to say, the use of most frequent content words – conversely to stopwords – seems to be enough to characterise an author's text as intended for children or adults. The words the authors use to target each of their age groups come easily to mind even when we think of similar texts for such age groups. For example, swear words ('fucking' or 'bloody [hell]'), and words that speak to adults, e.g., 'coffee' and 'whiskey', are present only in adult books. The same applies for 'sexual' and 'gay'. All authors make use of stronger and true-to-life words when writing for adult audiences, such as 'creed', 'anaesthesia, and 'police'.

When writing for minors, all authors, regardless of their specific topic of interest in their stories, use at least some fictional words, one would say straight out of fairy tales. We will stick to few examples, such as 'ginny' (most important feature for children's book in the all-authors model), 'clay' (Almond), 'imaginary' (Fine), 'dwarfs' and 'witch' (Gaiman), 'demons' (Pullman), and 'potion' (Rowling). We also spot words that appeal to children, for example, 'boring' and 'doll', and mentions of family members and relationships, especially 'uncle' and 'grandmother' which are present in the case of three authors.

6. Conclusion and future work

This study confirms that intended readership identification by only focusing on lexical features (content words, in particular) can be done in the case of the crosswriters examined. However, the lengthy nature of the segments, prevents us from performing a manual error analysis. Even so, as a whole, the models misclassified texts in a mere handful of cases, which proves that there is something distinctive in writing for either age group.

A potential indication that the author-specific models are not overfit-free yet is that the all-authors model has a significantly lower F1 macro score (.74 and .93 on average, respectively). To properly put an end to the overfitting issue, the robustness of each author's models should be tested, by feeding them with texts of other authors to observe more closely how they generalise.

As next steps to elaborate more on the subject, it would be interesting to examine what results an implementation with word embeddings in neural networks could yield. We have already seen the potential effectiveness of bidirectional long short-term memory (bi-LSTM) recurrent neural networks (RNNs) in NLP tasks (Li et al., 2019), as they are capable of not forgetting prior context by making use of their memory cells. In the case of RNNs, the danger of overfitting could also be further controlled by utilising dropout layers, i.e, by hiding a good portion (sometimes even as high as 50%) of the input that passes from one layer onto the next, so the model does not overlearn the data during training. In addition, the advent of pre-trained LMs has added to the already expanding interest the NLP field has gained over the last few years. By leveraging the billions of features that LMs are trained on, results that until recently would sound extraordinary to even the most optimists of researchers are now possible.

As far as this study is concerned, implementing the NER step and treating the entities as stopwords – rather than removing them from the original corpus altogether – has not been done before in the context of examining cross-topic authorship attribution, let alone the case of crosswriters. Not only do authors employ different words, but they also switch worlds in the process. They 'travel' from the real world to a fantasy one, and this is crystalised on their writing.

References

Argamon, S. & Shlomo, L. (2005). Measuring the Usefulness of Function Words for Authorship Attribution. *Proceedings of the Joint Conference on Association for Literary and Linguistic Computing/Association Computer Humanities*.

Beckett, S. L. (1999). *Transcending Boundaries: Writing for a Dual Audience of Children and Adults*. New York: Garland.

Berry, D.M. (2012). Introduction: Understanding the Digital Humanities. In: Berry, D.M. (eds) *Understanding Digital Humanities*. Palgrave Macmillan, London. doi:10.1057/9780230371934_1

Davani, A. M., Díaz, M., & Prabhakaran, V. (2022). Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics, 10*, 92–110. doi:10.1162/tacl_a_00449

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. doi:10.48550/arxiv.1810.04805

Finn, A. and Kushmerick, N. (2006), Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*, 57: 1506–1518. doi:10.1002/asi.20427

Gold, M. K. (Ed.). (2012). *Debates in the Digital Humanities*. University of Minnesota Press. http://www.jstor.org/stable/10.5749/j.ctttv8hq

Gomez Adorno, H., Rios, G. & Posadas D. J., Sidorov, G., & Sierra, G. (2018). Stylometry-based Approach for Detecting Writing Style Changes in Literary Texts. *Computación y Sistemas*, 22(1). doi:10.13053/cys-22-1-2882

Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques, *Literary and Linguistic Computing*, 22 (3), 251–270, doi:10.1093/llc/fqm020

Karsdorp, F., Kestemont, M., & Riddell, A. (2021). *Humanities Data Analysis: Case Studies with Python*. Princeton University Press.

Kestemont, M. (2014). Function Words in Authorship Attribution. From Black Magic to Theory?. In *Proceedings of the 3rd Workshop on Computational Linguistics for*

Literature (CLFL), 59–66. Association for Computational Linguistics. doi:10.3115/v1/W14-0908

Laramée, F. D. (2018). *Introduction to stylometry with Python*. Programming Historian. https://programminghistorian.org/en/lessons/introduction-to-stylometry-with-python.

Li, J., Xu, Y., Shi, H. (2019). Bidirectional LSTM with Hierarchical Attention for Text Classification. *IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 456–459. doi:10.1109/IAEAC47372.2019.8997969

López-Escobedo, F., Méndez-Cruz, C.-F., Sierra, G., & Solórzano-Soto, J. (2013). Analysis of Stylometric Variables in Long and Short Texts. *Procedia – Social and Behavioral Sciences*, 95, 604–611. doi:10.1016/j.sbspro.2013.10.688

Luyckx, K. & Daelemans, W. (2011). The effect of author set size and data size in authorship attribution. *LLC,* 26: 35–55. doi:10.1093/llc/fqq013

McIntyre, D. & Walker, B. (2019). *Corpus Stylistics: Theory and Practice*. Edinburgh: Edinburgh University Press.

Mikros, G. & Argiri, E. (2007). Investigating Topic Influence in Authorship Attribution. *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, PAN 2007*, Amsterdam, Netherlands, 276–282.

Sichel, H. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70, 542–547.

Stamatatos, E. (2009), A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60: 538–556. doi:10.1002/asi.21001

Stamatatos, E. (2007). Author Identification Using Imbalanced and Limited Training Texts. *Proceedings - International Workshop on Database and Expert Systems Applications, DEXA*. http://dx.doi.org/10.1109/DEXA.2007.5

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2001). Automatic Text Categorization in Terms of Genre and Author. *Computational Linguistics*, 26 (4): 471–495. doi:10.1162/089120100750105920

van Halteren, H. (2007). Author verification by linguistic profiling: an exploration of the parameter space. *Association for Computer Machinery Transactions on Speech and Language Processing*, 4(1): 1–17.

van Halteren, H., Baayen, H. R., Tweedie, F., Haverkort, M., and Neijt, A. (2005). New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12(1): 65–77.

Wu, H., Zhang, Z., & Wu, Q. (2021). Exploring Syntactic and Semantic Features for Authorship Attribution. *Applied Soft Computing*, 111(C). doi:10.1016/j.asoc.2021.107815

Zhao, Y., Zobel, J. (2005). Effective and Scalable Authorship Attribution Using Function Words. In Lee, G.G., Yamada, A., Meng, H., Myaeng, S.H. (eds) Information Retrieval Technology. AIRS 2005. *Lecture Notes in Computer Science*, vol. 3689. Springer, Berlin, Heidelberg. doi:10.1007/11562382_14

Zheng, W., & Jin, M. (2022). A review on authorship attribution in text mining. *WIREs Computational Statistics*, e1584. doi:10.1002/wics.1584

Appendix

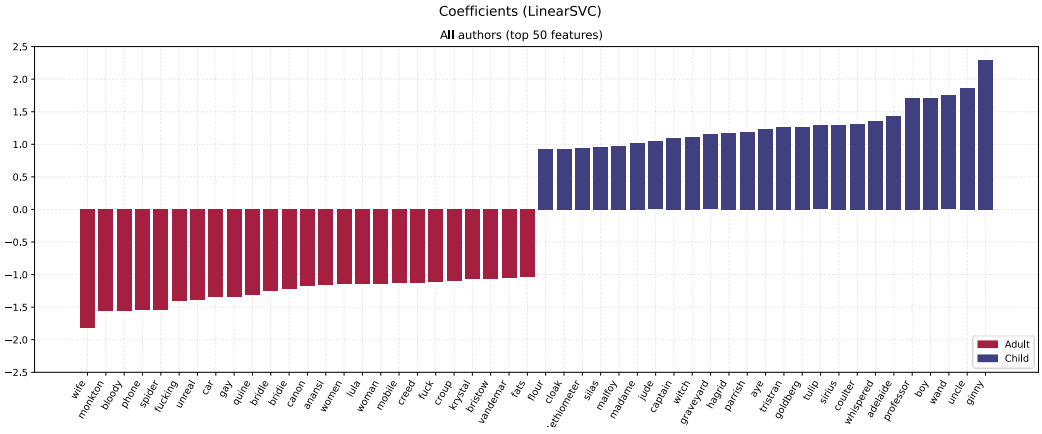

Figure 6: Plot of the feature coefficient of the Linear SVM model performed only on all texts. Children's books were selected as the positive class (blue), therefore the negative coefficients (red) indicate the 25 most contributing features for identifying books targeting adults.

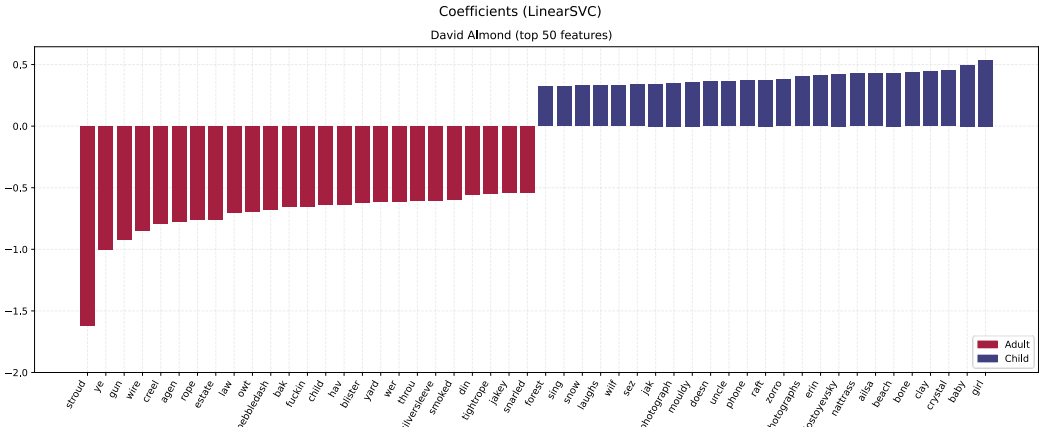

Figure 7: Plot of the feature coefficient of the Linear SVM model performed only on David Almond's texts. Children's books were selected as the positive class (blue), therefore the negative coefficients (red) indicate the 25 most contributing features for identifying books targeting adults.

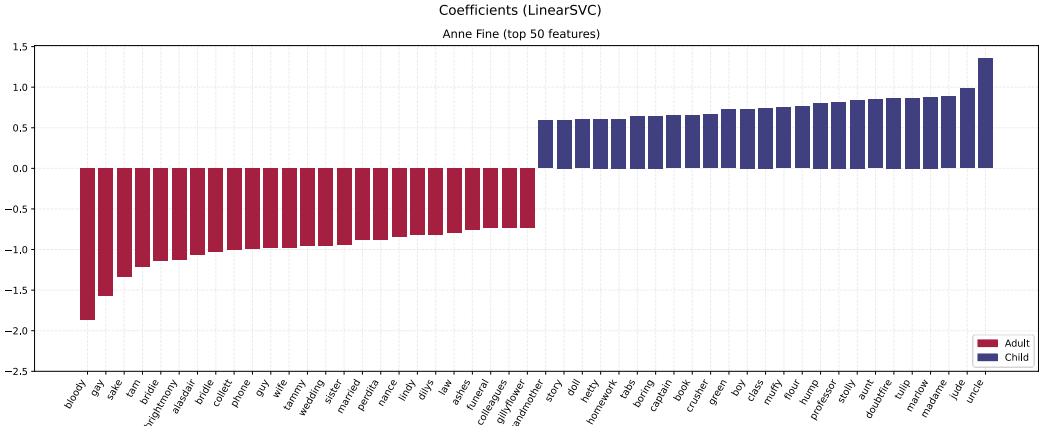

Figure 8: Plot of the feature coefficient of the Linear SVM model performed only on Anne Fine's texts. Children's books were selected as the positive class (blue), therefore the negative coefficients (red) indicate the 25 most contributing features for identifying books targeting adults.

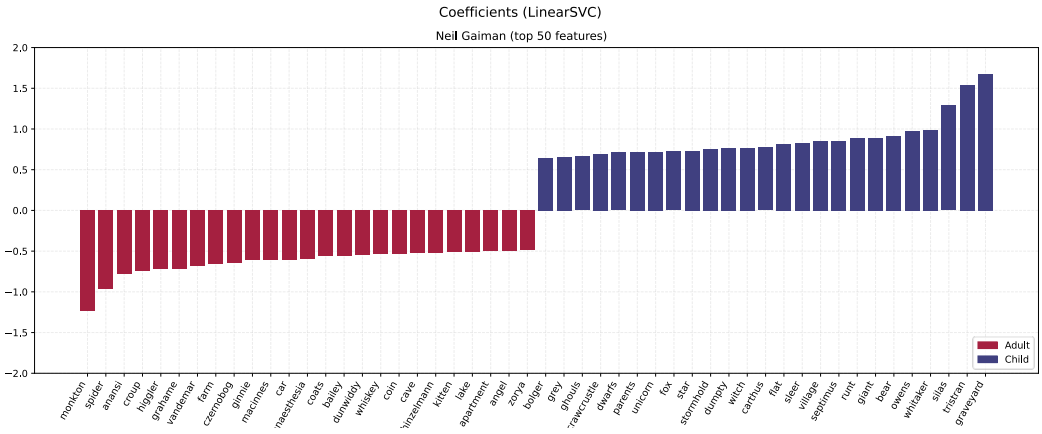

Figure 9: Plot of the feature coefficient of the Linear SVM model performed only on Neil Gaiman's texts. Children's books were selected as the positive class (blue), therefore the negative coefficients (red) indicate the 25 most contributing features for identifying books targeting adults.

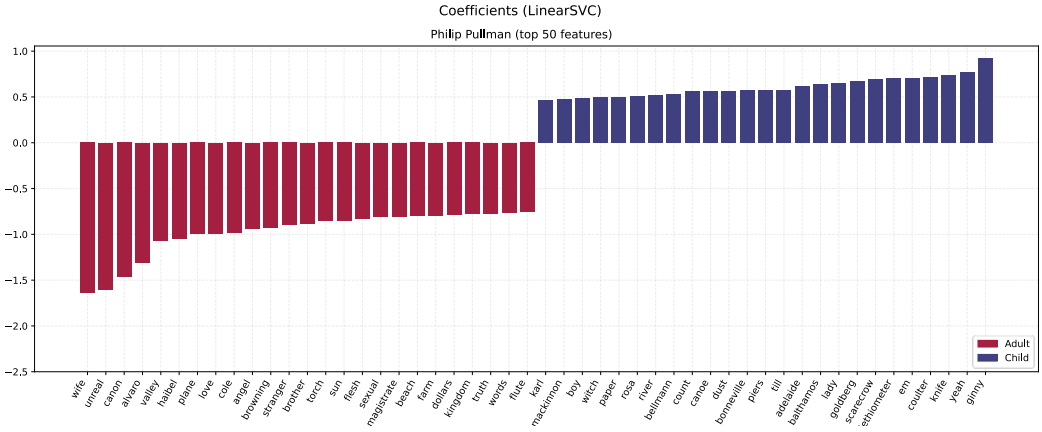

Figure 10: Plot of the feature coefficient of the Linear SVM model performed only on Philip Pullman texts. Children's books were selected as the positive class (blue), therefore the negative coefficients (red) indicate the 25 most contributing features for identifying books targeting adults.