Authorship Attribution: HuggingFace Keras Model with BERT Tokenization

Tessa de Vries tdevries@berkeley.edu
Shanie Hsieh shhsieh99@berkeley.edu
Shirley Jiang shirleyjiang@berkeley.edu

Abstract

Literacy is at an all time high with the firehose of written content available physically and virtually worldwide. It is more important than ever to be able to identify authors of given text, not only for rightful credibility but also intellectual property purposes. Our research applies text featurization techniques in an aim to match sample paragraphs to their proper author. In this paper we utilize a large breadth of literature stemming from several authors to address the classification problem of authorship attribution. In preliminary models, the breadth of author labels and the dearth of text examples presented a great challenge. However, after fine-tuning TensorFlow's HuggingFace Keras BERT base cased model, we were able to improve model accuracy and other evaluation metrics. Findings from our research contribute to the development of authorship attribution which lends positive implications for further historical studies, literary pursuits, and academic research.

Keywords: authorship attribution; plagiarism; TensorFlow; Keras; BERT tokenization

1 Introduction

The issue of authorship attribution becomes more and more challenging as the number of authors continues to grow at a fast pace. With a vast number of authors contributing to endless types of literature (e.g. academic papers, novels, poetry, and student writing), plagiarism and anonymous authorship identification become increasingly prevalent. Moreover, plagiarism and anonymous authorship identification stand as important literary issues that need addressing. However, this is quite challenging due to the breadth of authors and sparsity of examples.

Our goal is to build a model that identifies authors based on provided text samples to help build towards solutions to these aforementioned issues. We used a section of Project Gutenberg (Qian, He, et al., 2017), a public domain book archive, as the source of our data, which includes over 60,000 books and over 600 authors. Using a subset of Project Gutenberg, we built a data set from selected books, split these books into paragraphs and built a model that learns the writing style from these paragraphs with author labels to predict.

Based on previous research, we decided to fine-tune the Keras BERT (bert-based-cased) (Devlin, et al, 2019) transformer model hosted in TensorFlow's HuggingFace library (Wolf et al., 2020). A transformer model is fit to take in an input sequence and produce an output sequence. Our model learns from sentences in a paragraph where a transformer model best suits this. Keras is also one of the most used and user-friendly neural networks APIs that combine multiple back-end engines that best fits our resources and goal.

2 Background

Examples of authorship attribution include early examinations of *The Federalist Papers*, published under the pseudonym "Publius". While most of the 85 papers have been attributed to Alexander Hamilton, James Madison, or John Jay, twelve remained disputed. Mosteller's and Wallace's Bayesian processing using the IBM 7090 during the 1960's (Mosteller, Wallace, 1964) (Christopher, 2016) to Fung's SVM feature selection via concave minimization (Fung, 2003), are just a handful of text analysis tools that have been used to solve *The Federalist Papers* classification problem.

A more recent example of identifying anonymous authors is in the novel *The Cuckoo's Calling* by Robert Galbraith. It was revealed that "Robert Galbraith" is actually a pseudonym of *Harry Potter*'s J.K Rowling. In 2013, two researchers ran

the independent analyses of *The Cuckoo's Calling:* Peter Millican, using principal component analysis on word length, sentence length, paragraph length, letter frequency, punctuation frequency, and word usage features (Millican, 2003); and Patrick Juola, using 4-gram counts (Brooks, 2013) (Juola, 2015).

Furthermore, through machine learning applications on authorship attribution, linguists were able to find the likely authors of QAnon. (Kirkpatrick, 2022) One group of researchers used a proprietary software called OrphAnalytics, and the other used stylometry (Ainsworth and Juola, 2019).

What sets our research apart is that the above examples relied on prior knowledge to limit the pool of labels (authors). *The Federalist Papers* had all but twelve essays identified during the lifetimes of its original writers, allowing analysts to have labeled examples for each potential author. Analysis of *The Cuckoo's Calling* was predicated by an anonymous journalist's tip, prompting Millican and Juola to compare a small pool of works. Searches for QAnon authors looked at existing big names associated with QAnon. Our paper aims to have a generalized model trained on many examples of text and many author labels for authorship attribution.

Fabien, et al. (2020) fine-tuned a pre-trained BERT language model from HuggingFace for authorship attribution using the Enron Email, Blog Authorship, and IMDb62 data sets. Qian, et al. (2017) used the Reuters_50_50 data set and Project Gutenberg books with a Gated Recurrent Unit (GRU) network and a Long Short Term Memory (LSTM) network. We will combine and adapt this prior research by fine-tuning a pre-trained BERT language model from HuggingFace for authorship attribution using a generated data set of books from Project Gutenberg.

3 Methods

To reduce the massive range of authors and texts in order to have manageable data, we limited the data set to twenty authors with around ten respective books each. The books were accessed using the Gutenberg 0.8.2 API (Wolff, 2021) to source public domain books from Project Gutenberg which allows us to obtain raw text and text metadata. A limitation is that metadata lookup can only be completed when the metadata is locally cached. Running this code requires caching information for over 60,000 books. Using Google Colab GPU compute power, we were able to cache the metadata in approximately one hour.

Initially, we had selected 195 documents from twenty unique authors. Specifically, we hand selected authors with varying writer's styles and time periods (ranging from 4th century BC to the 1800s) to make the most diverse and rich training data. We first observed how our Keras model would perform in learning from the full documents. After feeding in these documents, we found we only had a 10% accuracy using the Adam optimizer at a 5e-5 learning rate, which is nearly equivalent to our model guessing one singular author for each given text. This accuracy, however, makes sense in the way the model is designed. The maximum input text length of BERT is 512 for single sentence classification Yang, et al., 2020, so the full document text that we were feeding in is too dense, making it so that our model is not learning how we expect it to. Additionally, there were not enough data points per label (author) for the model to completely grasp a writing style.

To combat this issue stemming from full document texts, we expanded the data set into paragraph level text by using double line breaks to define a paragraph. This exploded our data set from a relatively small size of 195 to an enormous size of 270,000+. It also became clear that within our data, different genres produced different types of "paragraphs". For example, a play consisting of almost entirely single line dialogue, has few sentences and words per "paragraph" as we defined.

At this point our data was too large to feed into our model, we ran into exhaustion errors without seeing training results. Thus we decided to take the 25th to 75th percentile (in word count) of paragraphs per book. Our thought process in doing so was to eliminate less useful paragraphs by removing the very short length ones. We noticed that some of our "paragraphs" were really titles of chapters, thus telling very little about the author's writing style. We also set a threshold of the 75th percentile as we noticed some paragraphs were excessively long and and could run into BERT's input maximum. Narrowing out data set to this range (the 25th to 75th percentile in length) reduced the size to around 140,182 data points.

Even after cutting our data set in roughly half, we still needed to reduce the size due to resource exhaustion errors. We deliberated the best way to do so, ensuring that we had a manageable amount

of data as well as preserved and accurately captured the style of our authors. We ended up taking 150 random samples from each author which left us with a very class-balanced data of 3000 data points.

Before hypertuning our models, we set a baseline model and accuracy for reference. Given we are solving a multi-class classification task, we set this baseline to be the majority class, meaning for each data point in the test set, our model predicts the majority class. In our case, our data was perfectly balanced thus the majority class could be any given singular author. This baseline model has an accuracy of 0.05.

After getting our data in its final form, we split it into training and testing sets and fine-tuned HuggingFace's Keras model hyperparameters. The Keras model is a pre-trained BERT language model with an added dense layer and softmax activation for authorship attribution (Wolf et al., 2020). Training was completed with Sparse Categorical Cross Entropy loss. We chose varied optimizers (Adam or Stochastic Gradient Descent), learning rates (5e-5 - 0.01), epoch numbers, the decay steps and the decay rate.

Our primary evaluation metric was accuracy. Not only was this built into the Keras model but accuracy is also widely accepted and used for balanced data sets. Specifically we used categorical sparse accuracy because as described in the Keras Metrics API, this metric is good for multi-class sparse text classification in deep learning. Given our problem statement of proper author identification, accuracy is the most important metric to build towards a solution above all other metrics.

4 Results and Discussion

As highlighted in Table 1, we ran several models with varying optimizers, number of epochs, learning rates, and decay steps and rates. The first six entries show different Adam optimizer based models. It is apparent that for these models, lower learning rates produced higher accuracies. More specifically, 5e-5 seemed to be the most optimal learning rate in conjunction with five epochs. From experimenting with varying epochs while holding the type of optimizer and learning rate constant, we found that five epochs was the most optimal. Lower than that, as shown with three epochs, produced lower accuracy likely due to underfitting. Conversely, higher than five epochs, shown with ten epochs, produced lower accuracy as well but

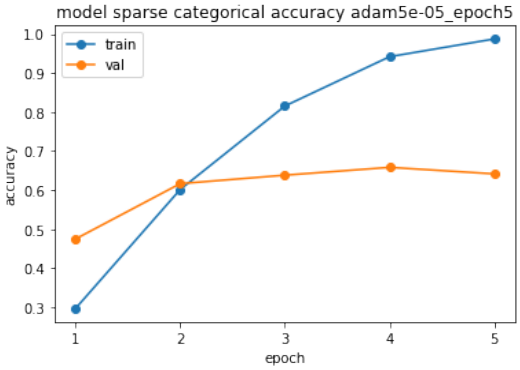

Figure 1: Best Model Accuracy

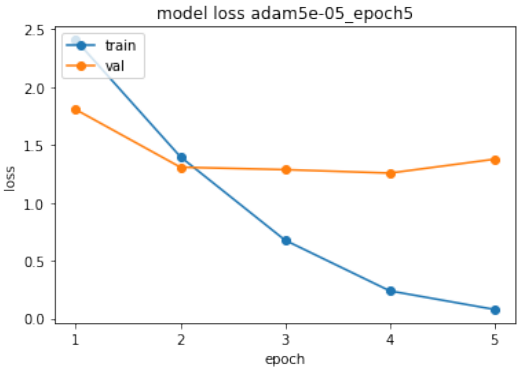

Figure 2: Best Model Loss

was likely due to overfitting. The lower half of the table shows our stochastic gradient descent (SGD) optimizer based models. These generally performed worse than the Adam optimizer based models. Amongst the SGD models, learning rate seemed to have less influence on performance than in the Adam based models, averaging at around 50% accuracy for the various SGD models. SGD is typically prone to underfitting perhaps due to the low learning rate since it learns on a small set of our data and builds on that.

Taking a dive deeper into our best performing model, we decided to analyze the sparse categorical accuracy and loss as the number of epochs progressed (from one to five). Looking at the left diagram in Figure 1, it appears that training accuracy gradually increases making its way to near 100%. Similarly, the validation accuracy gradually increases as well, however, at a much slower pace finishing around 65%.

Now taking a look at the model loss illustrated in the right diagram in Figure 2, we can see that the training and validation loss appear as inverse images to their respective model accuracies. While the training loss decreases at a rapid pace, ending

Optimizer	Learning Rate	Decay Steps	Decay Rate	Epochs	Val Accuracy
Adam	5e-5			3	0.6450
Adam	5e-5			5	0.6917
Adam	5e-5			10	0.6217
Adam	5e-2			3	0.0350
Adam	1e-5			3	0.5550
Adam	3e-5			3	0.4883
SGD	5e-5	10000	0.7	3	0.0583
SGD	5e-5	50000	0.9	3	0.0683
SGD	5e-5	10000	0.9	3	0.5417
SGD	1e-2	10000	0.9	3	0.5200
SGD	1e-2	10000	0.9	5	0.0883
SGD	1e-2	10000	0.9	10	0.0450

Table 1: Model Outcomes

around 0.0, validation loss decreases at a much slower rate. In fact, from epoch 4 to 5, there is a slight upward trend indicating the model may have overfit towards the end of training.

In the interest of confirming the accuracy we saw on the validation set, we created an additional test set composed of 200 unseen text samples from the same twenty authors. As expected, we got a very similar accuracy (68.5%) on this test set which reaffirmed our models ability to generalize well.

To further illuminate the best model's performance, we generated a confusion matrix (Figure A2) for a 200 row test set. Table A1 (found in the appendix) contains author metadata. It serves to help explain the model outputs.

In Figure A2, a handful of authors (Arthur Conan Doyle, J.M. Barrie, Mark Twain, Mary Shelley, WEB Du Bois, and L.M. Montgomery) have the lowest accuracy (60%). For example, Doyle is confused for Christie 20% of the time, which is not too surprising as both authors write about crime fiction. In fact, Christie was a fan of Doyle's early detective novels like Sherlock Holmes, so perhaps Doyle's writing inspired Christie's own style. Barrie, the author of Peter Pan and other books set in London, is confused for Dickens 20% of the time. Barrie wrote about fantasies for children in London and Dickens wrote about children coming of age in London, quite similar topics.

On the other hand, some authors like William Shakespeare and Plato had perfect predictions, this is likely due to their extremely distinct writing patterns. Shakespeare is the only playwright amongst the authors. Plato is separated from other authors by 2000 years and his works are the only translated works in the data set. The model was able to easily distinguish these authors.

While our model used similar data (Project Gutenberg) to research conducted by Qian, et al., we varied our model approach. Qian, et al. used a Gated Recurrent Unit (GRU) network and Long Short Term Memory (LSTM) network to achieve a very high accuracy near 99%. We intentionally pivoted away from these models and adopted the more modern Keras approach in order to explore a new solution space. Qian, et al. has already shown significant results, using the same data and models, there is not much more to improve on in terms of accuracy. With our highest performing Keras model, we achieved an accuracy around 69%, about 30% lower than the original paper. This may indicate that Keras is not the best suited model for this task; in future exploration, we would highly consider adopting a BERT based model in addition to our existing BERT based cased pre-trained transformer tokenizer.

5 Conclusion

Our model takes a step towards proper authorship attribution via deep learning. We hope it can serve as a preliminary model to other researchers or give insight into how to best approach this problem space. With the vast number of authors and even greater amount of existing literature spanning across numerous domains and time periods, we took a sliver to our model and fine-tuned it to achieve an accuracy of nearly 70%. We found using a Keras model with a BERT based cased tokenizer, Adam optimizer, learning rate of 5e-5 and 5 epochs, worked best given our specific and reduced data

set.

Our biggest limitation was not having a powerful enough machine with a larger GPU and RAM to digest and process a bigger data set. Given this, next steps would require a better machine to run our model on a larger data set that includes more text samples per author and perhaps even more authors. Including more authors could make our model even more generalizable. It would also be notable for future consideration to apply our model to a different data set containing other texts to assess the true functionality and accuracy. Moreover, to see if it can widely be in use for authorship attribution. In the future, we can explore other pre-trained BERT tokenizers or slacken BERT token limits to enable document-level attribution as opposed to our paragraph-level attribution. Ultimately, authorship attribution is a difficult task to tackle as the number of nameless and questionable authors of content appear each day, but our model attempts to close the gap and look for proper owners of text.

References

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship Attribution for Neural Text Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv:1706.03762 [cs]. ArXiv: 1706.03762.*

Ben Christopher., 2016. *How Statistics Solved a 175-Year-Old Mystery About Alexander Hamilton - Priceonomics.* Priceonomics.

David D. Kirkpatrick. 2022. *Who Is Behind QAnon? Linguistic Detectives Find Fingerprints*. The New York Times.

Chen Qian, Ting He, and Ren Zhang. "Deep Learning based Authorship Identification." (2017).

Claude-Alain Roten and Lionel Pousaz. 2022. *OrphAnalytics*, OrphAnalyticsSA.

Clemens Wolff. 2021. Gutenberg 0.8.2. PyPl

Frederick Mosteller, David Lee Wallace. 1964. Inference and Disputed Authorship: The Federalist.

Glenn Fung. 2003. The disputed federalist papers: SVM feature selection via concave minimization. In *TAPIA '03: Proceedings of the 2003 conference on Diversity in computing*, pages 42–46

Janet Ainsworth and Patrick Juola, *Who Wrote This?: Modern Forensic Authorship Analysis as a Model for Valid Forensic Science*, 96 Wash. U. L. Rev. 1161 (2019).

Liu Yang Mingyang Zhang Cheng Li Michael Bendersky Marc Najork. 2020. Beyond 512 Tokens: Siamese Multi-depth Transformer-based

Maël Fabien, 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)*, pages 127–137, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Noura Khalid Alhuqail. 2021. Authorship Identification Based on NLP. In European Journal of Computer Science and Information Technology, Vol.9, No.1, pp.1-26

Patrick Juola, The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions, *Digital Scholarship in the Humanities*, Volume 30, Issue suppl$_1$, $December 2015, $Pages i100–i113

Peter Millican. 2003. Signature1.0. PhiloComp.net

Richard Brooks and Cal Flyn. 2013. *JK Rowling, the cuckoo in crime novel nest*. Thetimes.co.uk.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, and Jamie Brew. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771 [cs]. ArXiv: 1910.03771.*

Yunita Sari, Mark Stevenson, and Andreas Vlachos. 2018. Topic or Style? Exploring the Most Useful Features for Authorship Attribution. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 343–353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Appendix

Table A1: Author Metadata

ID	Author Name	Years	Type	Genres
0	Shakespeare, William	(1564 - 1616)	Plays	Drama, Tragedy, History
1	Austen, Jane	(1775 - 1817)	Novel	Romance, Fiction
2	Dickens, Charles	(1812 - 1870)	Novel	Gothic romance
3	Thoreau, Henry David	(1817 - 1862)	Essay, Poem	Philosophy, Transcendentalist
4	Sinclair, Upton	(1878 - 1968)	Novel	Political fiction
5	Fitzgerald, F. Scott (Francis Scott)	(1896 - 1940)	Novel	modernist fiction
6	Doyle, Arthur Conan	(1859 - 1930)	Novel	Crime Fiction, Fantasy
7	Barrie, J. M. (James Matthew)	(1860 - 1937)	Novel, Plays	Children's literature, drama, fantasy
8	Twain, Mark	(1835 - 1910)	Novel	American Fiction
9	Shelley, Mary Wollstonecraft	(1797 - 1851)	Novel	Gothic romance
10	Plato	(~428BC - 348BC)	dialogue	political philosophy
11	Poe, Edgar Allan	(1809 - 1849)	Short Stories, Poem	Horror fiction, crime fiction
12	Baum, L. Frank (Lyman Frank)	(1856 -1919)	Novel	Children's literature
13	Du Bois, W. E. B. (William Edward Burghardt)	(1868 - 1963)	Essay	History, Sociology
14	Chesterton, G. K. (Gilbert Keith)	(1874 - 1936)	Essay, Novel	fantasy, Christian apologetics
15	Montgomery, L. M. (Lucy Maud)	(1874 - 1942)	Novel, Essay, Poem	Canadian literature, children's novels
16	Tapper, Thomas	(1864 - 1958)	Novel	Children's literature
17	Hawthorne, Nathaniel	(1804 - 1864)	Novel, Short Story	dark romanticism
18	Conrad, Joseph	(1857 - 1924)	Novel, Short Story	modernist fiction
19	Christie, Agatha	(1890 - 1976)	Novel, Short Story	Crime Fiction

Figure A2: Best Model Performance Confusion Matrix