Applying Transformer Contextualized Embeddings to Book of Mormon Authorship Attribution

Anonymous ACL submission

Abstract

Over the years, Latter-day Saint scholars have been interested in the possibility of empirically analyzing the purported authors of the Book of Mormon, even in the interest of providing evidence for multiple, separate authors of the text, as well as showing that the writing of the purported authors is measurably different than the writing of Joseph Smith. This previous work has relied on manually determined features (often referred to as *stylometry*). This paper attempts to utilize the contextualized embeddings that are the main benefit of Transformer models in order to use learned features in author classification and attribution.

1 Introduction

Latter-day Saint apologists and scholars alike have been interested in proving the validity of the faith's most unique and foundational text: The Book of Mormon. This sacred text, purportedly translated by Joseph Smith and published in 1830, is central to the tenets, practice, and doctrine of the Church of Jesus Christ of Latter-day Saints, as well as other denominations in the Latter-day Saint Restorationist movement.

This paper attempts to incorporate modern deep learning frameworks to the task of authorship attribution in the Book of Mormon, as an exercise. The end result is a visualization of the many sections of the text based on the purported author from the text, potentially demonstrating the plausibility of the text's own description of its origins.

2 Related Work

Previous attempts to use statistical methods to give strong empirical evidence of the non-Joseph Smith authorship of the Book of Mormon often centered around a methodology and set of assumptions called *stylometry*. The foundational paper to use this method is Larsen and Rencher (1982), but other iterations followed[1].

Book of Mormon Stylometry

One of the theoretical assumptions of stylometry is that there are certain identifiable aspects of one's own writing that are subconscious and consistent, and therefore not subject to conscious manipulation by the writer. Across the studies previously mentioned, analysis has often centered around the frequency and ratio of noncontextual words (e.g. *of*, *and*, *for*, etc.), and it has been posited that these ratios comprise a unique fingerprint, or *wordprint*, that will allow for the identification of the author of the text.

In their paper, Larsen and Rencher (1982) identified features based on noncontexual words, and then reduced those ratios using principle component analysis (PCA) in order to graph all the verses of the Book of Mormon based on these manual features. Additionally, they performed the same analysis on the known writings of Joseph Smith, and placed these datapoints on the same graph. Thus, their paper attempts to demonstrate that the Book of Mormon was not only written by multiple authors (as the text itself claims), but also that the stylometric features present in the Book of Mormon text are objectively and substantively different from the known writings of Joseph Smith. This is taken by believers as evidence for the divine origin of text, and a repudiation of claims that it was simply written by Joseph Smith himself.

Transformer Approaches to Authorship

In NLP terms, this previous approach to authorship attribution could be described as a classification problem that relies on manually determined features for that classification. However, machine learning and NLP applications have strongly

[1] See Roper et al. (2012) for a general overview of studies using stylometry in the context of Book of Mormon authorship.

demonstrated the ability of neural-based models to learn the necessary features and representations that are most useful for a given classification, instead of relying on manual feature engineering. This includes the potential use of the Transformer architecture (Vaswani et al., 2017), which has been so influential (even ubiquitous) in recent years. Indeed, Huertas-Tato et al. (2022) proposed a modified version of the Transformer framework that can be used specifically for authorship attribution.

In this paper, a more straight-foward method is used as an initial attempt at capturing the differences between purported authors.

3 Methods[2]

For this paper, a modified Transformer architecture was used, relying heavily on the ease and simplicity of use made available by HuggingFace[3], a framework and central hub that makes using Transformer-based models (among others) almost trivial. This particular paper utilizes a weaned-down version of the popular BERT Transformer, called DistillBERT (Sanh et al., 2019). This pretrained model is able to produce contextual embeddings for a given text, allowing the meaning of words to shift through vector space based on their context in a given text. This model is then fine-tuned for classification based on the specific corpus of the Book of Mormon.

The previous stylometric studies utilized the Book of Mormon text, with verses of the text tagged based on the purported author. Because the Book of Mormon claims to be a documentary text, which has been edited and compiled by multiple individuals, this provides the basis for the labelling. While this project did not have access to the exact same labelled corpus utilized by those initial papers, a database does exist that presents the text of the Book of Mormon with each author or speaker identified. Compiled by Hilton III and Hopkin (2012), each verse is broken up depending on the author, as indicated by the text itself (with ambiguous situations being decided on by the researchers who compiled it). These labels include other people who the authors appear to be quoting, and thus a single verse can be broken up into multiple voices.

This paper focuses solely on identifying the main authors of the text, and thus collapses any quoted text into being labeled with the author who did the quoting. This was accomplished by parsing the raw custom XML version of the database[4], and ensuring that each *voice section* (sometimes a verse, sometimes smaller based on voice switches) is labelled with the main author in question.

Next, this labelled data was used to fine-tune the DistillBERT model as explained above. As the previous research (Larsen and Rencher, 1982) concluded their paper with a diagram of the PCA-reduced features of each version, this paper attempts to do similarly. Instead of PCA, however, the more recent methodology *t-distributed Stochastic Neighbor Embedding* (or *t-SNE*) is used to generate a visualization of the features of each voice section (Van der Maaten and Hinton, 2008). After the model was fine-tuned, the final hidden layer of logits was used as an "authorship embedding", which was then reduced using t-SNE, resulting in a graph that shows the similarities and differences in two dimensions of each voice section.

4 Evaluation

4.1 Classification

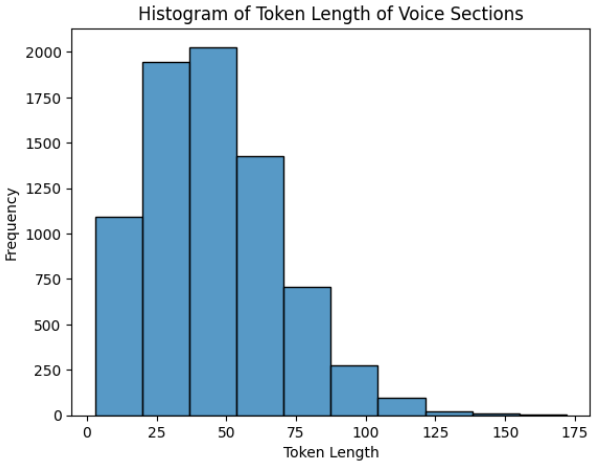

Figure 1: A histogram plotting the overall distribution of voice section length (by token).

In all, there were 7602 voice sections in the corpus. Figure 1 demonstrates how these sections are distribution by length, indicating that the mode number of tokens in each section is around 37.

[2]A GitHub repository with links to all requisite code can be found at https://github.com/reidempey/bomauthorship.

[3]The text classification code used to train the transformer for this paper drew heavily on the tutorial made available by HuggingFace found here: https://huggingface.co/docs/transformers/tasks/sequence_classification.

[4]Graciously provided to me by the BYU Department of Digital Humanities.

Author	Sections	Percentage	Accuracy
Mormon	4838	0.636	0.958
Nephi1	1690	0.222	0.793
Moroni2	733	0.096	0.411
Jacob	261	0.034	0.314
Enos	33	0.004	0.000
Amaleki	19	0.002	0.000
Jarom	16	0.002	0.000
Amaron	6	0.0007	0.000
Omni	3	0.0004	0.000
Abinadom	2	0.0002	0.000
Chemish	1	0.0001	0.000

Table 1: Number of Voice Sections by author, the percentage of the corpus those sections represent, and the accuracy of the classification by author.

As the model's ability to classify a text is likely affected by the text's length, Figure 2 indicates that the average accuracy increases as the texts become longer.

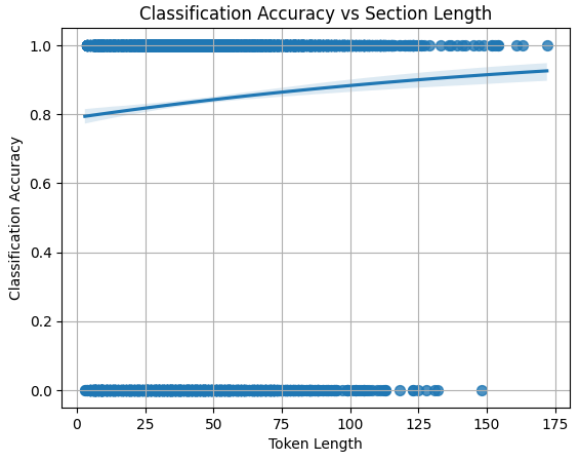

Figure 2: A graph plotting the successful or unsuccessful classification of each voice section as well as that section's length. This indicates that classification improves with section length.

Over the entire corpus, the overall accuracy[5] was 83.6%. As shown in Table 1, Mormon is the most prolific author of the Book of Mormon, representing over 60% of the text. Correspondingly, the accuracy of classifying the largest portion of the corpus is much higher at nearly 96%.

This begins to show the limitations of this paper's methodology. For the authors that have large contributions to the overall text, the classification accuracy tends to be much higher. However, as shown in Table 1, Only the top four authors were identified with any accuracy, and the remaining seven authors were not identified correctly a single time. Thus, this methodology is limited when it comes to unbalanced datasets.

[5]This accuracy refers to the measure commonly used in ML, which is calculated by dividing the true positives by all classification attempts.

4.2 t-SNE Dimensionality Reduction

Even with its limitations, the overall target of this paper is not simply classification, but to then use the classification to map a vector space with each voice section, and observe how well this captures a potential difference in authorship.

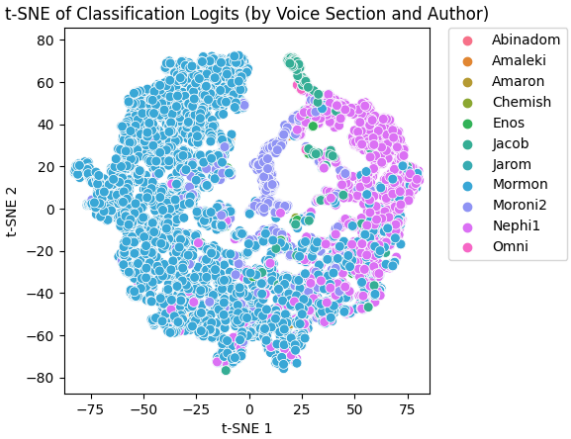

Figure 3: A t-SNE dimensionality reduction of the final hidden layer, marked by author.

As shown in Figure 3, there is clearly some overlap between even the major authors. Even so, the two most prolific authors (Mormon and Nephi1) generally appear to form separate clusters. Additionally, other authors seem to also have their own areas. Enos, for example, forms an almost linear cluser in between the main Mormon and Nephi1 clusters, perhaps suggesting that Enos' style is close to Nephi1's, but is still somewhere between the two.

5 Conclusions

This paper does not attempt to prove that the Book of Mormon was written by multiple authors, nor does it present any substantive innovations to improve the application of transformer architectures to authorship attribution. Rather, it attempts to apply a straightforward transformer architecture to the problem of Book of Mormon authorship attribution. There is no formal metric to determine if the t-SNE reduction in Figure 3 is good or bad, but

it is presented merely as food for thought about this type of research. There are many potential avenues for improving this methodology, as this paper was purposefully limited in scope for the purview of this class and for the sake of completion.

Future Work

Potential areas for future work include the following:

1. The original studies (incl. Larsen and Rencher (1982)) included the writings of Joseph Smith in order to demonstrate his own personal writing style to be different from that of the Book of Mormon. While evidence for multiple authors is already potentially confirmatory for believers, the inclusion of Joseph Smith's own writing would be enlightening.

2. The use of noncontextual words in previous work would help to prevent overfitting, which is possible in this current paper. It would be helpful to find a way to obscure some of the information that might be encouraging overfitting, while also not eliminating that information altogether. For this paper, a comparison was almost made to a corpus that simply included part-of-speech tags for contextual words, but as DistillBERT does not contain those POS tags in its vocabulary, this would amount to a simple mask. Other potential methods would be interesting to consider.

3. This paper used a simple transformer based on a HuggingFace tutorial. The insights of Huertas-Tato et al. (2022), and the use of their particular implementation of a fine-tuned transformer for authorship attribution would likely prove fruitful.

References

John Hilton III and Shon D Hopkin. 2012. Book of Mormon With Voices Identified.

Javier Huertas-Tato, Alvaro Huertas-Garcia, Alejandro Martin, and David Camacho. 2022. PART: Pre-trained Authorship Representation Transformer. ArXiv:2209.15373 [cs].

Wayne A Larsen and Alvin C Rencher. 1982. Who Wrote the Book of Mormon? An Analysis of Wordprints. *Book of Mormon Authorship: New Light on Ancient Origins*.

Matthew Roper, Paul J Fields, and G Bruce Schaalje. 2012. Stylometric analyses of the Book of Mormon: A short history. *Journal of Book of Mormon Studies*, 21(1):4.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.