CS 4650 Final Project: Neural Predictions of the Year of Authorship of Historical Texts

Everett Bolton and Christina Chen and Jonathan Gooch

Abstract

As media (and specifically literature) becomes more and more digitized, the need for accurate information regarding the origin of books (such as when they were written) becomes increasingly vital for researchers of the humanities. We aimed to create a Neural Network-based model that would predict the year that a given piece of text was written using only the text itself.

1 Introduction

With the increasing usage of digital systems to store media, the need to maintain accurate metadata can be key to both organization of digital archives and to the pursuit of humanities-focused research (especially research that is focused on specific authors or time-periods). However, the task of identifying the year of origin for a given book can be complex even when the author or publishing year of the book are known, as an author could be releasing old work significantly after it was initially written or could have their estate publish their work posthumously. In order to determine the year of origin for a given work without depending out potentially missing or unknown data such as authorship, we set out to create a model that would accept as input the text from said work and predict its year of origin using only the text itself (which should be present for any digital book even if no metadata related to it is).

2 Related Works

One related work comes to us from Stanford's CS224n: Natural Language Processing with Deep Learning class (Tausz, 2011). The project attempted to predict the year of authorship from the texts of books. One approach they utilized was to use the word-level n-grams as features for classifying the century of the book using Naive Bayes. The research shows that over 3 grams as features can drop the performance and that non-alphabetic words can have predictive power. In contrast to their approach of classifying the texts by century, our goal is to predict the exact year of origin of the text, which may be a more challenging task to achieve high accuracy for. The insight from this approach is that we use each year as a classification class in a discrete model and predict the label.

The paper also suggested an interesting approach of using a continuous linear model to calculate the year, where the prediction for the year is the summation of the feature counts multiplied by its least-square parameter and Gaussian noise. However, this model did not always perform well which suggested a non-linear relationship. The approach inspired us to use the normal distribution while labeling and build a continuous model for classifying the exact year.

The paper included other approaches and experiments that are worthy of consideration. For example, the low accuracy in the model could be improved by fitting in fewer classes of labels. The limitation of the paper is that the approaches of the paper are more algorithmic approach instead of neutral networks. In our approach, we look more into using machine learning.

3 Data Collection & Preparation

For our source data, we downloaded 5,051 English-language books from the Internet Archive using the Internet Archive API. Our source texts range in year from 1850 to 2023. After downloading, we reduced our dataset to exclude texts that were exceptionally small or repetitive (i.e., directories). Our dataset did not have an even distribution of publication years, so we decided to exclude books newer than 1950 to ensure that our models were trained on an even distribution. For each book, we excluded the first few pages of text because most books have the year of publication listed at the front. We chose to exclude these pages to ensure that our models are unable to "cheat" by reading the

publication year as an input token. We then built paragraphs from each book by splitting at double new-line characters, then excluded paragraphs that were unusually short or long. Each paragraph was tokenized with the appropriate tokenizer for a given model before training. The target data was either the integer value listed from the Internet Archive listing or a vector with a normal distribution centered around the target year (with semi-arbitrary stddev=5). This vector target approach was intended to provide some of the models with a target that allows some training reward for guessing years close to the target without treating this as a disjoint classification problem (where classes may not be considered interlinked). Additionally, this curved target data allows the model to predict a loose "era" of linguistic use, which more closely matches the underlying problem we are trying to solve.

4 Models

We took three approaches to different models to achieve our task.

First, we used DistilBERT and added a linear layer to convert a vector of probabilities for each year in the target data. Ideally this model will return values that are roughly a normal distribution curve where the peak of the curve occurs at the year of the source text. Our model uses the DistilBERT model to process the tokenized input text and then a linear layer maps the last hidden state vector (size 768) to a vector where the size is the number of years in the dataset (99 for 1850-1949).

Second, we used DistilBERT instead but mapped the output to a single fixed value to correspond to a target year. For example 0.0 would map to 1850 and 1.0 would map to 1949. This model has a somewhat simpler problem to solve as it has less weights to train, but also is unable to produce outputs that can describe the ambiguity of some input texts (where it may be nearly impossible to attribute a given text to a specific era, or could be from two different eras).

Third we made a model similar to the first, but instead used the I-BERT model in hopes that it would better handle longer input phrases (whole paragraphs as opposed to the sentences that BERT is trained on).

5 Results

Model	Train Loss	Valid Loss
1	$\mathbf{3.874 \times 10^{-4}}$	$\mathbf{4.798 \times 10^{-4}}$
2	3.926×10^{-2}	8.055×10^{-2}
3	4.332×10^{-4}	4.981×10^{-4}

Table 1: Losses for each model.

Model	Train Error	Valid Error
1	14.988	**23.349**
2	**14.869**	26.258
3	18.765	32.818

Table 2: Error margin in years for each model.

Model	Test Loss	Test Error
1	$\mathbf{5.3 \times 10^{-4}}$	32.684
2	0.14615	**32.273**
3	5.4×10^{-4}	37.608

Table 3: Test results for each model.

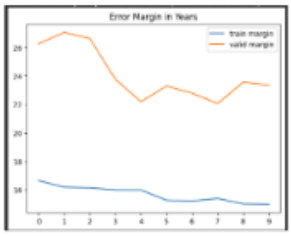
6 Discussion

Between the first and third model (which are of a similar structure) the first model appears to perform better on all metrics. The first and second models appear to perform similarly on the error margin metric, however the second model (which gives on floating-point return value) seems to be a little more accurate in validation. This was likely just because of slight differences in random values because the higher loss of the second model suggests that it is likely not more well generalized than the first model.

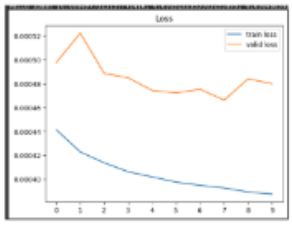
7 Conclusion

Our approach seems to be somewhat capable of approximating year of authorship for given paragraphs of text. Given models that could handle longer input texts (perhaps full books instead of paragraphs) we are confident that this approach would work more accurately. Additionally, models with larger and more complex encoders for full text (such as OpenAI's ada-002 model (Ryan Green, 2022)) could help the approach be more accurate.

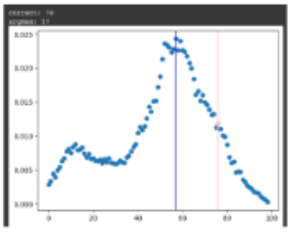
8 Images



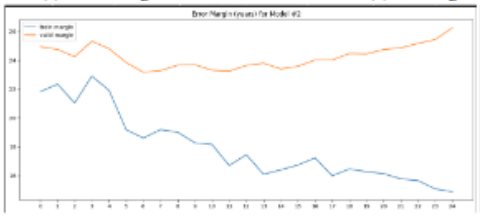
(a) Error Margin of Model 1



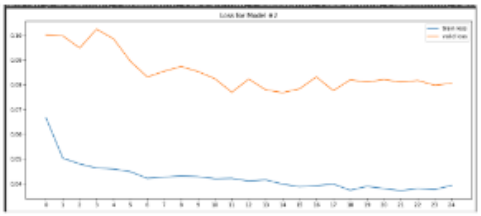
(b) Training Loss of Model 1



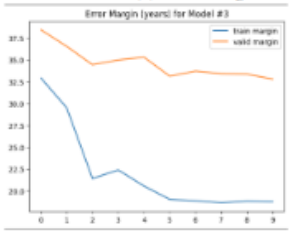
(c) Run of Model 1



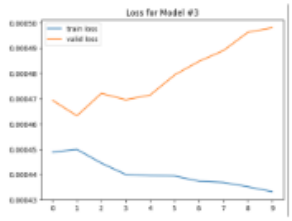
(d) Error Margin of Model 1



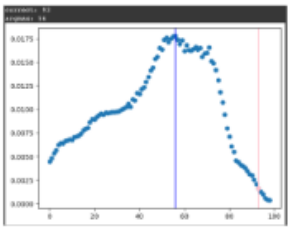
(e) Training Loss of Model 1



(f) Error Margin of Model 3



(g) Training Loss of Model 3



(h) Run of Model 3

References

Lilian Weng Arvind Neelakantan Ryan Green, Ted Sanders. 2022. New and improved embedding model.

Andrew Tausz. 2011. Predicting the date of authorship of historical texts.

(Tausz, 2011) (Ryan Green, 2022)