# Using Doc2Vec and Machine Learning to Predict Bible Verse Author and Genre

Julia Steed
*Min H. Kao Department of EECS*
*University of Tennessee*
Knoxville, United States
jsteed@vols.utk.edu

***Abstract*—Machine learning and natural language processing are rising in popularity for textual analysis. One field that has not been explored as much as others is how textual analysis can be applied to religious texts, such as the Bible. In this paper, we embed Bible verses using Doc2Vec [1]. Then, we train various logistic regression models to classify Bible verse genres and authors, comparing performance across models using different vector sizes. Additionally, we cluster the embedded verses and examine which genres and authors fall in the same clusters to identify similarities. This work can be used to gain insight into textual similarities between genres, which can help readers better understand how to interpret the Bible. Additionally, textual similarities across different authors may help to resolve authorship disputes. Lastly, this work showcases the capability of machine learning and word embedding to interpret religious text.**

***Index Terms*—Doc2Vec, natural language processing, text analysis, religion, machine learning, classification**

## I. Motivation

Machine learning (ML) and natural language processing (NLP) are rapidly growing technological domains. ML and NLP are used in virtually every field today including analysis of book text. However, one book that has not been studied in this way often is the Bible. There is some disagreement about authorship of certain books of the Bible, but it is generally established that somewhere between thirty and forty unique authors contributed to the Bible. The books of the Bible are split into two distinct sections, the Old Testament (OT) and the New Testament (NT), and represent roughly eight genres of literature, depending on how one defines the genres. As with any other book, the genre of Biblical text shifts how it should be read and interpreted; if the context is ignored, it can lead to misunderstanding and misapplication. Debates over how to interpret different genres have broken the church apart. For example, apocalyptic text primarily uses imagery and symbols that if taken literally can communicate an entirely different meaning. Premillennialists argue for a more literal interpretation unless the text suggests otherwise, while amillennialists interpret more figuratively. By using Doc2Vec to represent the verses of the Bible in ways a ML model can digest, I hope to analyze the similarities and differences between genres and evaluate whether ML is able to accurately predict a verse's genre based on features of the text. I plan to extend this project to also predict a verse's author and look at textual similarities and differences, which could be applied to authorship disputes.

## II. Dataset Overview and Exploratory Analysis

The data used in this project are sourced originally sourced from Kaggle [2] and GitHub [3]. The version of the Bible used for this project is the American Standard Version. However, for ease of use the data were uploaded to Demo Dataverse [4]. There are 8 unique genres and 29 unique authors represented in the data. Genres are distinctly separated by testament. The OT exclusively is made up of history, law, prophets and wisdom and the NT is made up exclusively of acts, epistles, apocalyptic, and gospels. I performed initial analysis to find the top 5 most common words in each genre. I found that the word "unto" is among the top five words for all 8 genres. Additionally, the Old Testament genres law, prophets, and wisdom have identical top five words, just in differing orders.

## III. Methodology

My first step was to sub-sample the original data by genre, ensuring even representation across verses. In order to leverage Doc2Vec to train a logistic regression model to predict a verse's genre, I generated a vocabulary of documents (verses) and tagged them with their corresponding genres. I built a Doc2Vec model with this vocabulary and trained it. Then, I generated vectors for each verse. I used these vectors as the features for my logistic regression model and the genres as labels. I performed hyperparameter tuning of the Doc2Vec vector_size parameter, testing 7 different values. I compared the resulting F1 scores and accuracies for each logistic regression model trained using the 7 different Doc2Vec models and kept the logistic regression model's parameters constant. I clustered the verse vectors produced by the best Doc2Vec model and used the elbow method to select the optimal number of clusters. I repeated this analysis to predict a verse's author for the Gospel authors.

## IV. Results

I found the optimal vector size for genre classification to be 100. Before sub-sampling the data, the logistic regression model achieved roughly 76 percent training and 73 percent testing accuracy. After sub-sampling, the model achieved around 92.5 percent training and 91 percent testing accuracy. The confusion matrix in Fig. 1 shows that the most commonly confused genres were acts with epistles, acts with gospels, acts with history, and prophets with wisdom.

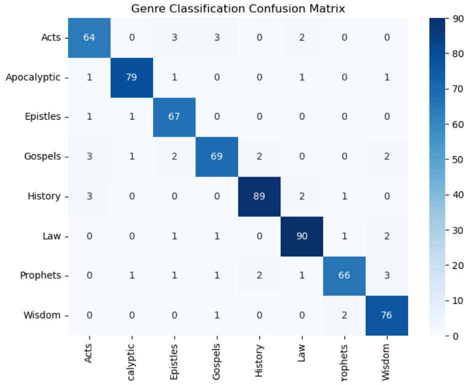

Fig. 1. Genre Classification Confusion Matrix

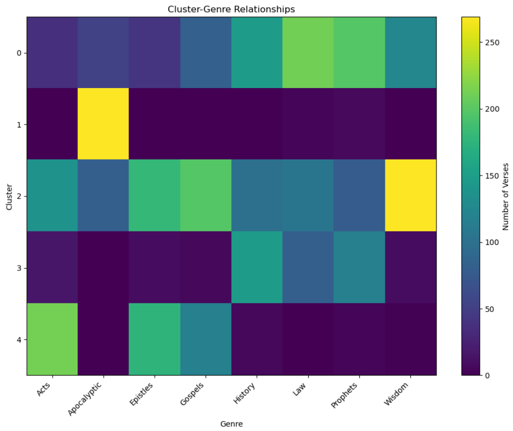

Fig. 2. Cluster-Genre Relationships

I found the optimal number of clusters for genre classification to be 5. Fig. 2 shows the placement of genres within clusters. Clusters 0 and 3 reveal that history, law and prophets have similar features. Clusters 4 and 2 show that acts, gospels, and epistles are also similar and have some commonalities with wisdom. Clusters 1 and 2 show that apocalyptic and wisdom are the most distinct genres because they dominate these two clusters.

I attempted to sub-sample authors as well, but found that there were not enough data points to do this effectively. When predicting on the full dataset, the model achieved around 67 percent training and 64 percent testing accuracy. To obtain more meaningful results, I decided to use a subset of the data that only contained verses written by the Gospel authors.

I found the optimal vector size for author classification to be 50. The logistic regression model achieved around 79 percent training and testing accuracy. The confusion matrix shows that some of the most commonly confused authors were Matthew with Mark and Luke with Mark.

I found the optimal number of clusters for genre classification to be 3. Fig. 2 shows the placement of authors within clusters. Mark and Luke dominate cluster 2, while Matthew and Mark dominate cluster 0.

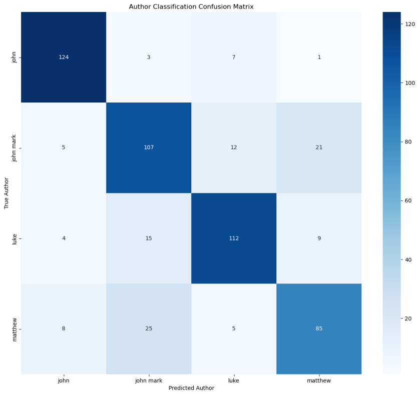

Fig. 3. Author Classification Confusion Matrix

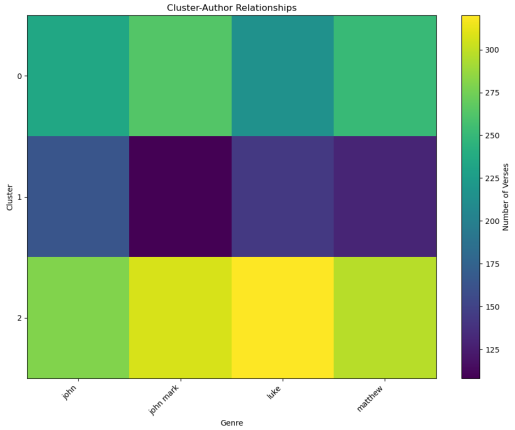

Fig. 4. Cluster-Author Relationships

## V. Future Work

This project could be expanded to predict many other data points in the Bible, such as predicting whether a verse is part of a parable or predicting whether a verse was spoken by a specific person, such as Jesus. Additionally, one could repeat this analysis for different translations of the Bible to see whether the results differ substantially. One could also apply the author analysis to other other select genres with more distinct authors, such as prophets.

## VI. References

### References

[1] "gensim: topic modelling for humans," Radimrehurek.com, 2013. https://radimrehurek.com/gensim/models/doc2vec.html

[2] "Bible Corpus," www.kaggle.com. https://www.kaggle.com/datasets/oswinrh/bible (accessed Apr. 20, 2024).

[3] jinwook, "happygrammer/bible-metadata," GitHub, Jun. 13, 2021. https://github.com/happygrammer/bible-metadata/tree/main (accessed Apr. 20, 2024).

[4] "Demo Dataverse," demo.dataverse.org. https://demo.dataverse.org/ (accessed Apr. 20, 2024).