Predicting Bible Verse Genre and Author

Julia Steed

Overview

Project Goal: Leverage text embedding and machine learning (ML) to predict a Bible verse's genre and author.

Motivation

- Rise in popularity of ML and NLP for textual analysis
- Religious texts have not been explored as often as other texts
- Genre affects the way text should be read and interpreted
- Ignoring context given by genre can lead to misinterpretation, which can lead to divisions in the church
- Authorship disputes for many books of the Bible

Dataset

- Originally from Kaggle and GitHub, uploaded to Dataverse
- 8 genres: Acts, Apocalyptic, Epistles, Gospel, History, Law, Prophets, Wisdom
- First 4 are exclusively NT, last 4 OT
- 29 unique authors + category for "Unknown"
- Around 31000 verses total
- "unto" is among the top 5 words for each genre
- Law, Wisdom and Prophets all have identical top 5 words (unto, thy, shall, Jehovah, thou), in different orders

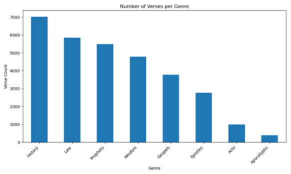


Figure 1: Verses per Genre

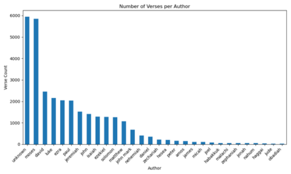


Figure 2: Verses per Author

Exploratory Data Analysis

- My first step was performing initial EDA to identify potential issues when applying ML techniques
- Genres and authors are unevenly distributed (Fig 1 & 2)
- Mitigate for genre model by subsampling
- Mitigate for author model by filtering to Gospel authors

Methodology

Models

- First, I generated a vocabulary of documents (verses), tagged by their corresponding label (genre or author)
- I built a Doc2Vec model, trained it and generated verse embedding vectors to be used as the features for Logistic Regression
- I performed hyperparameter tuning for Doc2Vec vector_size to find best model
- Analyzed misclassifications
- I clustered the vectors resulting from the best model and analyzed cluster-genre and cluster-author relationships

Results

Genre Classification

- Optimal vector_size was 100 (Fig. 3)
- 92.5% training and 91% testing accuracy
- Most commonly confused genres were Acts with Epistles, Acts with Gospels, Acts with History, and Prophets with Wisdom (Fig. 4)

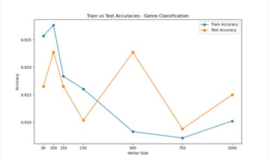
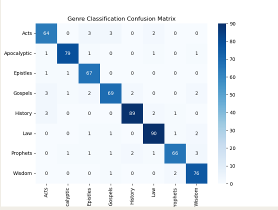


Figure 3: Genre Accuracy



	Acts	calyptic	Epistles	Gospels	History	Law	rophets	Wisdom
Acts	64	0	3	3	0	2	0	0
Apocalyptic	1	79	1	0	0	1	0	1
Epistles	1	1	67	0	0	0	0	0
Gospels	3	1	2	69	2	0	0	2
History	3	0	0	0	89	2	1	0
Law	0	0	1	1	0	90	1	2
Prophets	0	1	1	1	2	1	66	3
Wisdom	0	0	0	1	0	0	2	76

Figure 4: Genre Confusion Matrix

- Optimal number of clusters I found using KMeans was 5 (Fig. 5)
- Clusters 0 and 3 show that Law and Prophets have similar textual features
- Clusters 2 and 4 show that Acts, Epistles, and Gospels have similar textual features
- Clusters 2 and 3 show that Apocalyptic and Wisdom have the most distinct features

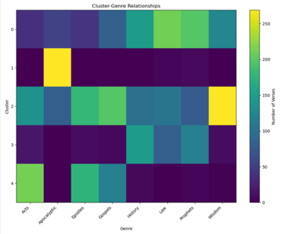


Figure 5: Cluster-Genre Heatmap

Results

Author Classification

- Optimal vector_size was 50 (Fig. 6)
- ~79% training and testing accuracy
- Most commonly confused authors were Matthew with Mark and Luke with Mark (Fig. 7)

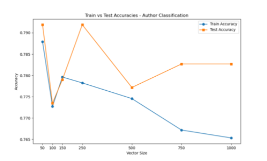
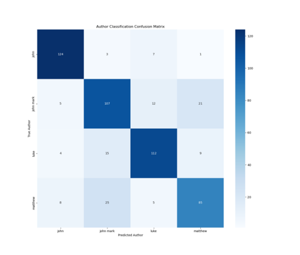


Figure 6: Author Accuracy



	john	john mark	luke	matthew
john	124	3	7	1
john mark	5	107	12	21
luke	4	15	112	9
matthew	8	25	5	85

Figure 7: Author Confusion Matrix

- Optimal number of clusters I found using KMeans was 3 (Fig. 8)
- Cluster 0 is dominated by Matthew and Mark
- Cluster 2 is dominated by Mark and Luke

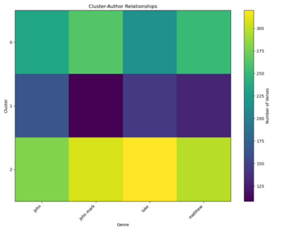


Figure 8: Cluster-Author Heatmap

Future Work

- This project could be expanded to predict other data in the Bible, such as whether a verse is part of a parable or not, or whether a verse was spoken by a particular person, like Jesus
- Additionally, this analysis could be repeated with different translations of the Bible to see if the results are substantially different from the ASV
- Lastly, one could apply the author analysis to a more diverse genre, such as the Prophets

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE