Issues in Information Systems
Volume 23, Issue 1, pp. 181-189, 2022

DOI: https://doi.org/10.48009/1_iis_2022_114

The authorship of hebrews: A topic modeling analysis

George Ray, *Shepherd University, gray@shepherd.edu*

Abstract

An increasing arsenal of computational tools is employed to enrich the research done in the humanities. One of those tools, Latent Dirichlet Allocation (LDA) is used in this research to identify the latent topics in the New Testament once it is collated into two corpora: Paul's books; and the other books in the New Testament. The purpose is to investigate a longstanding theological question concerning Paul's authorship of the New Testament Book of Hebrews. The hidden topics discovered by LDA are mapped through posterior analysis to the books in the New Testament to determine their Paulene content. The findings from this analysis do not support Paul's authorship of Hebrews. The latent topics that generated Hebrews are more closely related to the other authors than to Paul. Perplexity analysis is used to determine the optimal number of topics.

Keywords: Textual Analysis, Artificial Intelligence, Digital Humanities, Latent Dirichlet Allocation.

Introduction

Digital humanities projects are digitizing large collections of archival information, which permits researchers to connect with classic works in new ways. Among the many collections digitized is the New Testament. In addition to textual criticism and linguistic analysis, now digital methods can be used in its analysis. A digital transformation is occurring in the humanities and computational methods are now being applied to analyze research questions.

Analysis of unstructured data with a complex organization is done based on several approaches: 1.) the statistical characteristics of the natural language text; 2.) network modeling for analysis of complex linguistics; 3.) longitudinal analysis of language; and 4.) the nature of syntax. The digitization and analysis of literary content has provided useful insights in the humanities (Gerlach & Font-Clos, 2019; Kolaczyk & Czardi, 2014).

The main objective of this study is to test the hypothesis that Paul is the author of Hebrews using digital methods. The methodology used is Topic Modeling, which is a textual analytic approach in the digital humanities that uses machine learning to search for natural groupings in a corpus. A key assumption behind Topic Modeling is that each word in a document is selected by the author in two steps: first, the selection of topics to address in the book or document; and second, the selection of the words and phrases the author is familiar with to express each topic. The purpose of Topic Modeling is to discover the latent structure of topics that generated a corpus of documents (Blei, 2012).

Topic Modeling

This research uses Topic Modeling to classify the New Testament Book of Hebrews as to authorship by finding the latent topic structure of Hebrews and comparing the topics that Paul uses in his books with the topics in Hebrews. Our assumption is that the topics to be found in Paulene books are distinct and we can use the proportion of those topics found in any book to classify it as being authored by Paul. Topic Modeling will reveal the proportion of words in a book that are generated from a topic and with that the researcher can estimate if certain topics are strongly associated with certain books.

Latent Dirichlet Allocation (LDA) is an approach to Topic Modeling that estimates the mixture of words associated with a topic and the mixture of topics associated with a document (Silge & Robinson, 2017). It is based on the distributional hypothesis that similar topics use similar words and the mixing hypothesis that a document has a mixture of topics with an associated statistical distribution.

There are two types of probability distribution considered in LDA: 1.) P(w|t) or the probability distribution of seeing a word given a topic; and 2.) P(t|d) the probability distribution of seeing a topic given a document (Silge & Robinson, 2017; Grus, 2015). The second type, P(t|d) will be used as a discriminator in this article. The idea is to group New Testament books based on the hidden structure of topics in the corpus, compare the topics used to generate Hebrews with those found in books known to be authored by Paul.

Other analytic methods are used to classify observations. Vehkalahti and Everitt (2019) discuss the application of cluster analysis as a tool of classification. In an early example of digital tools being applied to questions in humanities, Roaf (1983) used cluster analysis to answer a question posed by archaeologists and art historians about the number of sculptors at the Apadana in Persepolis.

Atkinson-Abutridy (2020) finds that the Topic Modeling ability to find hidden structures is an advantage it has over traditional clustering methods. The generative process in Topic Modeling helps uncover associations in a corpus. LDA is a generative model that is useful in classification. Mimno and McCallum (2007) were able to use LDA with 3 large books to effectively form a recommender that finds related books "that share the same topic, organized in descending order by the proportion of the book assigned to the topic" (p. 382).

Literature Review on Hebrews

The Book of Hebrews is written as an encouragement to believers (American Revision Committee, 1901, Heb 13:22). Other epistles in the New Testament owe their topical expression to Hebrews, including Peter's epistles (Forster,1838). In his inquiry for the Archbishop of Canterbury, Forster (1838) also analyzed the literary style and structure of Paul and the apostles. His conclusion is that Peter's works were influenced by Paul (p. 627).

The anonymous Book of Hebrews was first attributed to Paul over a hundred years after it was written. Other authors have been suggested including Luke, Barnabas, and Silvanus. Manson (1951) finds compelling stylistic similarities between Hebrews and Stephen's speech in Acts. Paulene authorship would lend additional weight to the book with Paul learning scriptures under the guidance of Gamaliel, one of the renowned Rabbinic scholars, forming the concepts and expressions for his later works in a formal setting (Robbins, 1861). The arguments for authorship over the years have been based on external testimony as well as internal proofs.

The King James Version attributes authorship to Paul, while Calvin and Luther reject Paulene authorship. By and large, however, modern western criticism rejects Paul as the author of Hebrews noting that while the writing style has both similarities and differences with Paul's other writings, the differences in the sophistication of the Greek are significant. In contrast, the eastern church has consistently advanced Paul as author of the text (Mitchell, 2007, p.2). This study fills a gap in the literature by applying digital methods to investigate this question. This is part of the digital transformation in the humanities.

Gutenberg Project

Access to large humanities datasets is available through the Gutenberg Project, Wikisource and other text digitization projects. The Gutenberg Project has a major advantage in digital humanities research in that it is a large, cataloged set of literary works that is open source. Project Gutenberg is a volunteer effort started in 1971 that has grown into an online digital library with more than 60,000 volumes that have been uploaded by the community (Gutenberg Project, 2022). Interested parties can access the library through its web site using search tools to find books to read online or download. Books come with metadata that contain bibliographic information on the work.

The New Testament is available in the Gutenberg Project. In this research, it is accessed and read into an R program that is used to apply natural language computations (Ray, 2022). The *gutenbergr* code library in R has a *gutenberg_download* function that takes the Gutenberg ID of a book as a parameter and returns the book as a data frame containing the text.

The text processing functions of R can then be applied. The Gutenberg identification numbers of the New Testament books are 8347 through 8373. The downloaded New Testament books are collected into a corpus, which is then composed into a document term matrix.

The twenty-seven books in the New Testament deal with similar concepts and have a common context. It is safe to assume that any differences in topics are due to the choices made by authors rather than the genre of one set of books versus that of another set. Every book is treated as a document and is a mixture of topics. The extent a document uses words and phrases from a particular topic associates that document with that topic.

Methodology

The research question is whether Paul wrote the Book of Hebrews, and the research will classify Hebrews based on the Paulene topical content of that book. An approach is developed that uses LDA as its foundation to examine the topical similarity between a document in question with Paul's corpus in the New Testament. Would the Paul's corpus reveal a hidden structure that could generate the document in question from its words and phrases? Topic Modeling is used to discover which books in the New Testament gather into Paulene books and other author books in such a way that authorship of Hebrews can be evaluated.

In this study, the R programming language is used to collect the books of Paul into one corpus and books by the other authors in the New Testament into a second corpus. The two are merged into an NT corpus with a document field added to show "Paul" and "other" appropriately. A document term matrix is then formed and, with the appropriate number of topics, an LDA model produced.

After this processing on the main corpus, the next step is to convert the target book alone into a document term matrix. This matrix is then run through the posterior algorithm along with the topic model just generated. LDA is also used to evaluate each book of the New Testament as to the proportion of content

derived from the Paulene topics and the proportion from the 'other' topic. This is likewise done using the posterior analysis function for the LDA model.

The results returned from the posterior analysis include a distribution of the target book over the model's topics; in other words, what percentage of the book is associated with each topic. The individual proportions of the various Paulene topics in the intersection are aggregated into a percentage of topics in the target book's topic distribution that also generate Paulene books. Then Hebrews can be compared using the proportion of it derived from Paul with each of the other books in the New Testament and their proportions derived from the Paulene topics.

Perplexity

The New Testament document term matrix is fed into the LDA function available from the *topicmodel*s code library in R. In addition to the document term matrix, the number of topics, known as K, must be included as a parameter. The perplexity measure of a resulting topic model can be used to compare the effectiveness of different values for K. *topicmodels* has a function *perplexity* that calculates the perplexity for a model. This scoring can be used to compare models. The model having the lowest perplexity score is considered the best (University of Chicago, 2021). Blei, Ng and Jordan (2003, p. 1008) also hold that a lower perplexity score indicates better performance. The perplexity score is used in this study to determine the number of topics.

Gan and Qi (2021) find that while perplexity may evaluate the effectiveness of topic training, it often results in large number of topics. The consequence of this is topics that are similar may appear, causing poor recognition of topics. They propose that Jensen-Shannon Divergence, "a measure of the difference in word-probability distributions between topics" (p. 5), be used to verify that the topics generated are well differentiated. They recommend using an index comprised of perplexity and other measures to improve the effectiveness of determining the number of topics.

Glowacka-Musial (2022) applied Topic Modeling to automate the cataloging of subject headings at New Mexico State University Library. She developed a performance metric, recall, which is the fraction of subject headings assigned manually by a cataloger that are present in the list of candidate headings produced by the topic model. She compared this recall metric with the results of a perplexity analysis and both methods recommended the same optimal number of topics.

Validation

The last aspect of methodology is validation. Leave-one-out cross-validation (LOOCV) is a useful testing process for small datasets. The object is to compare predictions made by a model against actual values (Rad & Maleki, 2020). In this case, one book is used as a testing set while the other books generate the model. The model is then applied to the book in the training set and proportions noted. This is repeated with each of the books being used as the test dataset. The Book of Hebrews is not included in this validation as its authorship is anonymous and it is the subject of this research. Zhang and Wang (2016) consider LOOCV to be an effective testing process.

Forster (1838) found that Paul influenced other New Testament writers but did not find an influence of the other writers on Paul. Considering this, the books authored by Paul should show a high content of Paul's topics. On the other hand, the books of the other authors should show a varying balance between Paul's topics and the 'other' topics.

Results

Perplexity Calculation

A processing loop was established to compute the topic model for a different number of topics varying from two through fifty and derive the perplexity for each such model. The results are shown in Figure 1 with the lowest perplexity score occurring at three before starting to rise to higher levels. A lower perplexity score indicates a more effective number of topics. A K of three for the corpus of Paul's books and other author's books in the New Testament is an effective number of topics.

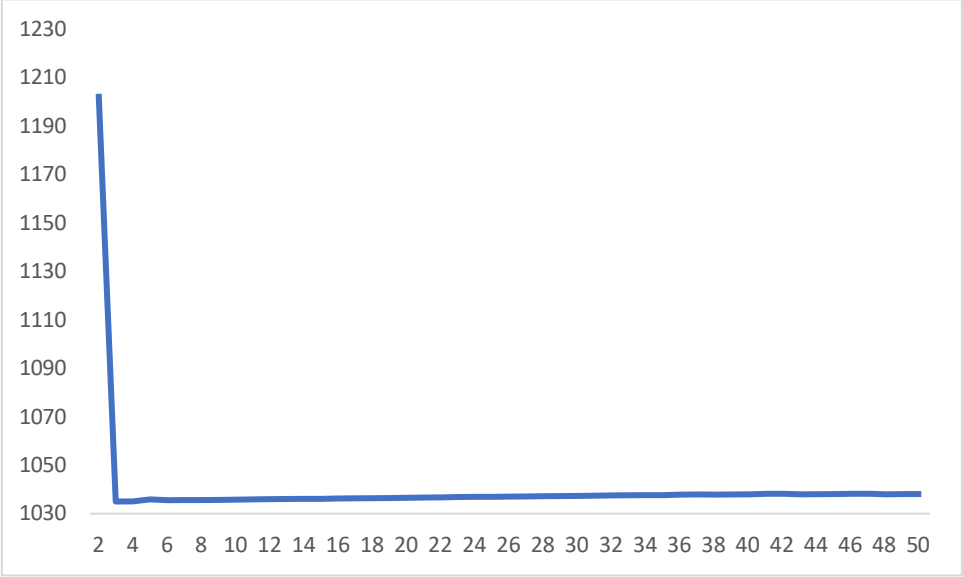

Figure 1: Perplexity distribution

Authorship of Hebrews

The topic model returned by the LDA contains several matrices, including the gamma matrix that has the per-document-per-topic probabilities. These are aggregated into a structure that has the topics associated with each document. Next the target book, and each book in turn is the target book, uses the topic model for a posterior analysis that shows the topics associated with that book. For each book, the percentage of Paulene topics in its topical structure is shown in Table 1.

Table 1: Proportion of Content from Paulene Topics

Target	Proportion	Target	Proportion
Matthew	0.00%	1 Tim	99.99%
Mark	0.00%	2 Tim	99.98%
Luke	0.00%	Titus	99.97%
John	0.00%	Philemon	99.94%
Acts	0.00%	Hebrews	45.80%
Romans	100.00%	James	34.14%
1 Cor	100.00%	1 Peter	55.64%
2 Cor	99.99%	2 Peter	42.30%
Galatians	99.99%	1 John	38.69%
Ephesians	99.99%	2 John	57.18%
Philippians	99.98%	3 John	52.49%
Colossians	99.98%	Jude	34.05%
1 Thes	99.98%	Revelations	0.00%
2 Thes	99.97%		

These results are graphed in Figure 2.

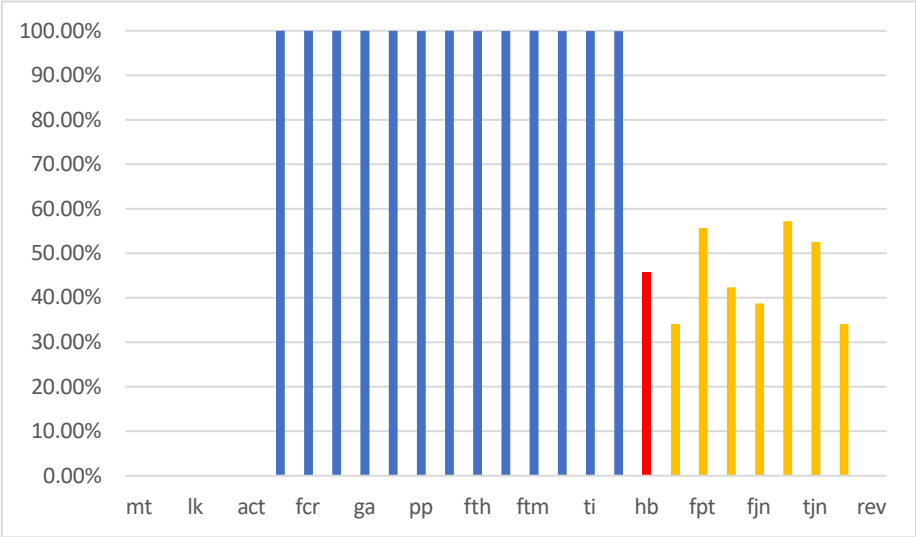

Fig. 2: Content Generation of New Testament Books from Latent Paulene Topics

Forty five percent (45%) of the Book of Hebrews is generated from topics that were also used to generate the Paulene texts. The books with Paul as the known author show that a high percentage of the text is generated by the Paulene topics from the LDA model. These books, Romans through Philemon, are in the center of the chart and their range is from 99.94% to 100%.

Low percentages of the Gospels, Acts and the Book of Revelation are effectively 0%. The non-Paulene epistles, starting with Hebrews in red and continuing to the right in gold, show a varying adoption of Paulene topics, but they do not come close to the proportions shown by the works known to be authored by Paul. The range of Paulene topic proportion found in the non-Paulene epistles ranged from 34% to 57%. The Epistle to the Hebrews is in the middle of that range. These findings do not support Paul as the author of

Hebrews but, like the other epistles in this set, Hebrews does borrow from Paul's works.

Discussion

The results for the Paulene books are much as would be expected. An average of 99.8% of the content in Paul's books are generated by the topics closely associated with Paul. The average of the books not authored by Paul is 25.74%. The book of Hebrews comes in at 45.8%. The results do not support the conclusion that the topics that generate Hebrews are those that generate the Paulene books. First Peter shows 55.64% of its content generated from Paulene topics.

The approach used in this study produced results that are as expected for the books with known authors. The LDA analysis in this study finds no support for the hypothesis that the topics that generate Hebrews are those that generate the Paulene books. This provides an element of support for the modern western opinion on the authorship of Hebrews.

The contribution of this paper is the application of digital technology to theological analysis to create new processes for theological research. This is part of the digital transformation in the humanities. Another objective of the study is to develop a framework to discover and surface insights from disparate, unstructured data sources. The framework developed for this research applied digital methods to classify a book from antiquity. The framework is modular with modules for accessing the source data, preprocessing the data, organizing it according to the research needs, a module for applying LDA to classify, and a module to display results.

A major limitation of the framework used for this research is that it is currently capable of classification but not for digital studies on influence, the evolution of a topic over time, and correlation of topics. Digital methods for Document Influence Models (Shalit, Weinshall & Chechik , 2013), Dynamic Topic Models (Blei & Lafferty, 2006) and Correlated Topic Models (Blei & Lafferty, 2007) have been developed but are not currently available in the framework.

A practical implication of this study is that it builds on the digital library initiatives, such as the Gutenberg Project. These projects have expanded our ability to digitize scholarly resources. To explore ideas using those resources, studies like this paper are needed to apply digital methods to digitized resources. We can now apply both quantitative and qualitative research methods to a wide variety of cultural texts. There are many existing qualitative studies to provide context for questions being investigated.

Another implication of this study is it began development of a modular framework to do analysis on the now available digital libraries. This framework can substitute various tools to tailor the analytic process to a particular set of data from the wide variety of source categories in the humanities. This study tested a straightforward hypothesis by building a set of modules to access, organize and analyze with the purpose of classifying one of the books as to author.

Other questions such as topic evolution over time, correlation among topics and topic influence on other topics require more framework development and similar studies. For this purpose, the programming code has been made available on GitHub. Others may add to this foundation to create a general-purpose tool that can be used in the digital transformation of the humanities.

No financial support was received for this study or paper.

References

American Revision Committee (1901). The Holy Bible. Prolific Industries, LLC. Troy, Michigan.

Atkinson-Abutridy, J (2020). Text Analytics. Atkinson-Abutridy.

Blei, D. (April 2012). Probabilistic Topic Models. *Communications of the ACM, vol. 55, no. 4*

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3(4/5), 993–1022*. https://dl.acm.org/doi/10.5555/944919.944937

Blei, D. & J. Lafferty (2006). Dynamic Topic Models. *Proceedings of the 23rd International Conference on Machine Learning.*

Blei, D. & J. Lafferty (April 2007). A Correlated Topic Model of Science. *The Annals of Applied Statistics, 2007, Vol. 1, No. 1, 17–35*

Forster, C (1838). The Apostolical Authority of the Epistle to the Hebrews. Published by James Duncan, 37. Paternoster - Row, London. Reprinted by BiblioLife.

Gan, J. & Y. Qi, (Oct 2021) , Selection of the Optimal Number of Topics for LDA Topic Model—Taking Patent Policy Analysis as an Example. *Entropy, Vol. 23 Issue 10, p1301-1301*.

Gerlach, M. & F. Font-Clos (November 29, 2019). A Standardized Project Gutenberg Corpus for Statistical Analysis of Natural Language and Quantitative Linguistics. *Entropy, 22, 126; doi:10.3390/e22010126*

Glowacka-Musial, M. (June 2022). Applying Topic Modeling for Automated Creation of Descriptive Metadata for Digital Collections. *Information Technology & Libraries, Vol. 41 Issue 2, p1-14*.

Grus, J. (2015). Data Science from Scratch. O'Reilly.

Gutenberg Project (2022). Welcome to Project Gutenberg. Retrieved from https://www.gutenberg.org/

Kolaczyk, E. & G. Csardi. (2014). Statistical Analysis of Network Data with R. Springer.

Manson, W. (1951). The Epistle to the Hebrews: A Historical and Theological Reconsideration. Hodder And Stoughton, London.

Mimno, David and McCallum, Andrew. (2007). Organizing the oca: learning faceted subjects from a library of digital books. *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, pp. 376–385.* ACM.

Mitchell, A (2007). Hebrews. Liturgical Press.

Rad, K. & A. Maleki (Sep 2020). A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology). Vol. 82 Issue 4, p965-996. 32p*.

Ray, G. (2022). topicLDA: a Digital Humanities Application in R. Available at https://github.com/gmrwvu/DigitalHumanities

Roaf, M. (1983). Sculptures and sculptors at Persepolis. IRAN, Jouranl of the British Institute of Persian Studies 21, 1-159.

Robbins, R. D. C. (1861). Was The Apostle Paul the Author of The Epistle to the Hebrews? *Bibliotheca Sacra & Biblical Repository. 1861 3rd Qtr, Vol. 18 Issue 71, p469-535. 67p*. Retrieved from Ebscohost.

Shalit, U., D. Weinshall & G. Chechik (2013). Modeling Musical Influence with Topic Models. *Proceedings of the 30 th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR:W&CP volume 28*.

Silge, S & D. Robinson (2017). Text Mining with R. O'Reilly.

Vehkalahti, K. & B. Everitt (2019). Multivariate Analysis for the Behavioral Sciences, 2e. CRC Press.

Weizhong, Z., J. Chen, R. Perkins, L. Zhichao, G. Weigong, D. Yijun & Z. Wen (2015). A heuristic approach to determine an appropriate number of topics in Topic Modeling. *BMC Bioinformatics. 2015 Suppl 13, Vol. 16, p1-10. 10p.*

Zhang, J. & S. Wang (Aug 2016) A fast leave-one-out cross-validation for SVM-like family. *Neural Computing & Applications, Vol. 27 Issue 6, p1717-1730. 14p.*