Using Latent Dirichlet Allocation to Summerize the Bible

Alena McCain

***Abstract*— Latent Dirichlet allocation is a reasonably reliable unsupervised learning technique for determining the topics in various text data. However, LDA hasn't been used widely on one of the bestselling books in the world, the Bible. My goal is to summarize the Bible using LDA. To help me achieve this, I first determine which translations of the Bible should be input for the LDA models. I then create multiple LDA models for each Bible translation, determine their optimal parameters, and compare the models that use the optimal parameters. While performing LDA on individual books of the Bible has been attempted, I couldn't find any literature on LDA performed on the entire Bible. With this in mind, the results of this paper intend to be an insightful starting point for future research.**

Keywords— Latent Dirichlet Allocation, Bible, coherence

I. Introduction

The Bible is the best-selling book of all time, with over 5 billion copies printed, according to MPR news [1]. However, a survey completed in 2014 by the Pew Research Center [2] states that 45% of Americans rarely or never read the Bible. In addition, 68% of Americans who rarely or never read the Bible believe in God. Therefore, people might be more interested in reading the Bible or growing in the Christian faith in general if it was easier to understand the critical topics in the Bible.

Garbhapu et al. [3] compared the performance of LDA and LSA (Latent Semantic Analysis) by dividing the book of Genesis into roughly 70 documents and determined how similar each document was to all the other documents. They also compared the most common words found by LDA and LSA and each algorithm's coherence score.

Similarly, Hu [4] performed LDA on Psalms and Proverbs. They separated both books into documents by chapter and then by verse. Since Psalms was primarily written by David and Proverbs was written by Solomon, they wanted to see if there were any similarities between the ideas written by a father and his son. The results of this research were primarily to cluster the documents in Psalms and Proverbs. It did not discuss the topics that the LDA model produced in detail.

Although the research topics discussed above used LDA on the Bible, neither performed LDA on the whole Bible. They were also more focused on comparing verses and chapters in the Bible based on topic modeling clusters instead of discussing the words in the topics produced by the LDA models. I provide topic modeling and exploration over the entire Bible, analyze whether LDA has practical applications for topic modeling in the Bible, and suggest parameter estimates for LDA algorithms to aid future work in this area. I perform topic modeling on multiple translations of the Bible to explore if and how various translations of the Bible affect LDA's performance and topic selection.

The rest of this paper is arranged as follows. Section II lays out the data that is used throughout the experiments discussed in this paper. It also discusses initial observations regarding the data being used for natural language processing. Section III presents the selection methods and results of determining which Bible translations to use on LDA. Section IV outlines the document makeup on which the LDA models were executed and the preprocessing done on the data before the LDA models are created. Section V discusses the methods and results of parameter selection for the LDA models. Section VI explains the LDA experiments that were conducted and their results. Lastly, the conclusion and potential future work are presented in Section VII.

II. Data Description and Exploration

A. Data Description

The data [5] used throughout this paper for experimentation are the individual verses of the Bible, organized by Bible translation. The data includes all the verses of the American Standard Version, Bible in Basic English, Darby English Bible, King James Version, Webster's Bible, World English Bible, and Young's Literal translations of the Bible. All of the verses are in English.

B. Data Exploration Related to Natural Language Processing

Many names, places, and other essential ideas in the Bible are not common in the English language as we use it today. Therefore, before performing any natural language processing on the Bible, I needed to determine how many of these uncommon words are included in Python's Spacy vocabulary because that could affect the results presented later in this paper.

To accomplish this, I referenced "Biblical Names and their Meanings" by Brad Haugard [6]. According to the author, "It contains more than 2,500 Bible and Bible-related proper names and their meanings." I determined that 73% of the names in this document were not in Spacy's vocabulary. While this might seem like a large number, I determined that this would not affect my experimentation very much. Many of the names in that 73% do not appear frequently enough in the Bible to belong to a topic.

III. Bible Translation Selection

A. Selection Criteria

In 1991, William Chamberlain [7] published the "Catalog of English Bible Translations." While the book does contain some duplicate and blank entries, it is still a massive 900 pages long, listing roughly 900 English Bible translations.

With this in mind, I wanted to find the most representative translation of the English Bible, so LDA models wouldn't have to be created for every translation. Here representative is defined as being most similar to other translations based solely on the content in the verses (the words themselves). For comparison, the least representative translation was selected for future comparison between the topic selection and functionality of LDA on both translations.

B. Methodology

The following steps were taken to determine which Bible translations are the most and least representative. First, every verse was labeled by translation. Next, all of the identical verses were removed across the different translations. For example, the first verse of Genesis is written the same in the World English Bible and the Darby English Bible. Therefore, that verse was removed in both translations because it would be impossible to classify correctly. A few different text classifiers were then trained on the remaining verses to see how well they were classified by translation. The translation with the smallest number of correctly classified verses is considered the most representative. The translation with the largest number of correctly classified verses is considered the least representative translation.

C. Experiments and Results

As stated in the previous subsection, identical verses among two or more translations of the Bible were removed before using a machine learning classifier. After doing so, I determined that roughly 86% of all verses across all translations were unique. I also looked at how many unique verses there were for each translation individually, relative to the total number of verses in the Bible. These results are shown in Table I. According to the table, Kings James and Webster's Bible translations are potentially the most representative, simply because they share the most identical Bible verses with the other translations.

Before determining which translations are the least and most representative, it is essential to analyze overall discrepancies in the F1 scores of the models shown in Figures 1-4. In every case, it would appear that the Bible in basic English, Word English Bible, and Young's Literal Translation generally have higher F1 scores compared to the rest of the models. Part of this discrepancy could be that the models with lower F1 scores also have a lower percentage of unique verses. Therefore, the models have fewer verses to train on, leading to lower precision. However, this is inconsistent with the findings of the Darby English Bible because the Darby English Bible has a large percentage of unique verses but a low F1 score. This discrepancy could be because some Bible translations are semantically similar, meaning they are harder to classify. This can be seen when comparing figures 2 and 4 because removing the Kings James Version increased precision in the other translations. The recall for all translations increased in Figure 4 because removing one translation allowed for more test verses of the other translations. In Figures 1-3, some translations appear to have a low recall causing lower F1 scores, even if they have a high percentage of unique verses (like the Darby English Bible).

TABLE I. THE PERCENT OF VERSES UNIQUE TO EACH TRANSLATION

Translation	***Percent of Unique Verses***
American Standard Version	85.16
Bible in Basic English	97.33
Darby English Bible	93.84
King James Version	64.86
Webster’s Bible	67.81
World English Bible	93.13
Young's Literal Translation	98.33

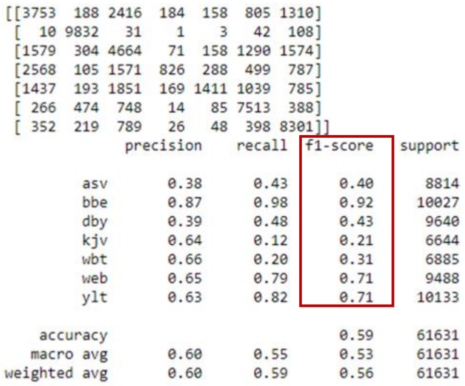
With this information in mind, I proceeded to run three different classification models on all of the Bible verse data, discarding the identical verses, to see which translation's verses were the least and most correctly classified. Figures 1-3 show the results of this using a Naïve Bayesian, Linear SVC, and Logistic Regression classifier on the Bible verse data. When running the Linear SVC classifier, a thousand iterations were executed. When running the Logistic Regression classifier, 100 iterations were executed. All of these parameters were the default parameters for these models. I decided against doing any parameter tuning because these classifiers were used to get a general idea of the differences between the Bible translations and is not the primary purpose of this paper. In all three cases, the test set was 33% of the entire data set.

Regarding the Naïve Bayes classifier in Figure 1, the Kings James Version had a noticeably low F1 score and recall. Regarding the Linear SVC and Logistic Regression classifiers, the King James Version had the second-lowest metrics next to the American Standard Version. Given this information and the fact that the King James Version has a much higher percentage of duplicate verses compared to the American Standard Version, I decided the King James Version would be most representative.

Next, to verify my findings and determine the least representative Bible translation, I decided to run a linear SVC classifier using every translation except the King James Version. I decided to use linear SVC because that model had the highest accuracy and F1 scores (for the most translations) of the previous classifiers tested. This classifier was executed on the same conditions as the previous Linear SVC classifier, and the test set was 33% of the entire data set.

Figure 4 shows the results of running this classifier on every Bible translation except the King James Version. As one can see, the accuracy increased from 64% in Figure 2 to 70%, and the F1 scores improved at least slightly for every translation, except the Bible in Basic English. It should be noted that the American Standard Version had an F1 score increase of 13%, which is a dramatic increase. This reinforced my belief that the King James Version is the most representative translation.

Even though the F1 score did not increase for the Bible in Basic English between Figure 2 and Figure 4, that isn't significant because that translation consistently outperformed the other translations in every model. This observation led me to conclude that the Bible in Basic English is the least representative of all the translations.

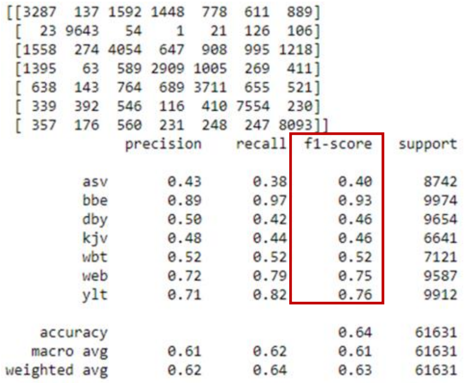


```
[[3753  188 2416  184  158  805 1310]
 [  10 9832   31    1    3   42  108]
 [1579  304 4664   71  158 1290 1574]
 [2568  105 1571  826  288  499  787]
 [1437  193 1851  169 1411 1039  785]
 [ 266  474  748   14   85 7513  388]
 [ 352  219  789   26   48  398 8301]]
              precision    recall  f1-score   support

         asv       0.38      0.43      0.40      8814
         bbe       0.87      0.98      0.92     10027
         dby       0.39      0.48      0.43      9640
         kjv       0.64      0.12      0.21      6644
         wbt       0.66      0.20      0.31      6885
         web       0.65      0.79      0.71      9488
         ylt       0.63      0.82      0.71     10133

    accuracy                           0.59     61631
   macro avg       0.60      0.55      0.53     61631
weighted avg       0.60      0.59      0.56     61631
```

Fig. 1. The confusion matrix and corresponding metrics of Bible verse classification by translation using the Naive Bayes classifier. The translations are as follows; asv: American Standard Version, bbe: Bible in Basic English, dby: Darby English Bible, kjv: King James Version, wbt: Webster's Bible, web: World English Bible, ylt: Young's Literal Translation.



```
[[3287  137 1592 1448  778  611  889]
 [  23 9643   54    1   21  126  106]
 [1558  274 4054  647  908  995 1218]
 [1395   63  589 2909 1005  269  411]
 [ 638  143  764  689 3711  655  521]
 [ 339  392  546  116  410 7554  230]
 [ 357  176  560  231  248  247 8093]]
              precision    recall  f1-score   support

         asv       0.43      0.38      0.40      8742
         bbe       0.89      0.97      0.93      9974
         dby       0.50      0.42      0.46      9654
         kjv       0.48      0.44      0.46      6641
         wbt       0.52      0.52      0.52      7121
         web       0.72      0.79      0.75      9587
         ylt       0.71      0.82      0.76      9912

    accuracy                           0.64     61631
   macro avg       0.61      0.62      0.61     61631
weighted avg       0.62      0.64      0.63     61631
```

Fig. 2. The confusion matrix and corresponding metrics of Bible verse classification by translation using the Linear SVC classifier The translations are as follows; asv: American Standard Version, bbe: Bible in Basic English, dby: Darby English Bible, kjv: King James Version, wbt: Webster's Bible, web: World English Bible, ylt: Young's Literal Translation.

IV. Document Makeup and Preprocessing

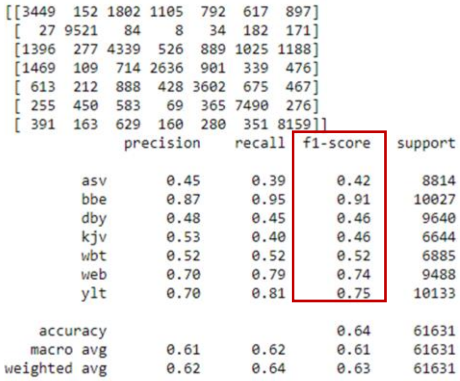
A. *Document Makeup*

As determined in section II, LDA models were executed on the King James and the Bible in Basic English translations. There were two different document makeups; each document was a book of the Bible (66 total documents), and each document was a chapter of the Bible (1189 total documents).

B. *Preprocessing*

In the LDA models presented in the rest of this paper, the following preprocessing steps were taken; every word was converted to lowercase, punctuation was removed, stop words were removed, and every word was lemmatized if possible. It should be noted that a word was only lemmatized if it was a noun, adjective, verb, or adverb.

In addition to the standard stop words, I also included a subset of words from The King James Bible Online [8]. That list includes some of the most frequently used words in the King James translation of the Bible. I decided to use this list because most words in the Bible probably are not included in a standard stop word list. Leaving those words could cause the models to focus on frequent but unimportant words. In doing preliminary testing, I determined that using these additional stop words increased the coherence values of my potential models by roughly 10%. However, it should be noted that I only ran a few tests for each translation and document makeup. Removing these extra stop words never decreased the coherence values relative to the original stop words, so I was reasonably confident that they would be helpful in preprocessing.

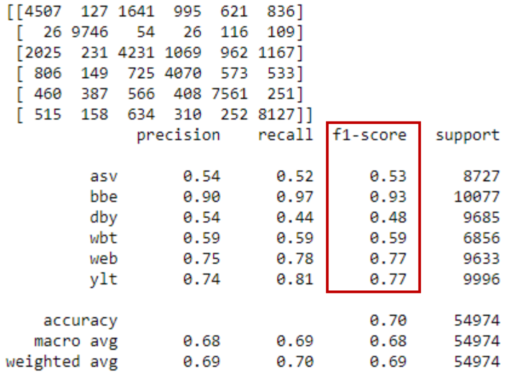


```
[[3449  152 1802 1105  792  617  897]
 [  27 9521   84    8   34  182  171]
 [1396  277 4339  526  889 1025 1188]
 [1469  109  714 2636  901  339  476]
 [ 613  212  888  428 3602  675  467]
 [ 255  450  583   69  365 7490  276]
 [ 391  163  629  160  280  351 8159]]
              precision    recall  f1-score   support

         asv       0.45      0.39      0.42      8814
         bbe       0.87      0.95      0.91     10027
         dby       0.48      0.45      0.46      9640
         kjv       0.53      0.40      0.46      6644
         wbt       0.52      0.52      0.52      6885
         web       0.70      0.79      0.74      9488
         ylt       0.70      0.81      0.75     10133

    accuracy                           0.64     61631
   macro avg       0.61      0.62      0.61     61631
weighted avg       0.62      0.64      0.63     61631
```

Fig. 3. The confusion matrix and corresponding metrics of Bible verse classification by translation using the Logistic Regression classifier The translations are as follows; asv: American Standard Version, bbe: Bible in Basic English, dby: Darby English Bible, kjv: King James Version, wbt: Webster's Bible, web: World English Bible, ylt: Young's Literal Translation.



```
[[4507  127 1641  995  621  836]
 [  26 9746   54   26  116  109]
 [2025  231 4231 1069  962 1167]
 [ 806  149  725 4070  573  533]
 [ 460  387  566  408 7561  251]
 [ 515  158  634  310  252 8127]]
              precision    recall  f1-score   support

         asv       0.54      0.52      0.53      8727
         bbe       0.90      0.97      0.93     10077
         dby       0.54      0.44      0.48      9685
         wbt       0.59      0.59      0.59      6856
         web       0.75      0.78      0.77      9633
         ylt       0.74      0.81      0.77      9996

    accuracy                           0.70     54974
   macro avg       0.68      0.69      0.68     54974
weighted avg       0.69      0.70      0.69     54974
```

Fig. 4. The confusion matrix and corresponding metrics of Bible verse classification by translation using the Linear SVC classifier without the King James Version The translations are as follows; asv: American Standard Version, bbe: Bible in Basic English, dby: Darby English Bible, wbt: Webster's Bible, web: World English Bible, ylt: Young's Literal Translation.

V. LDA Parameter Selection

A. *Parameter Selection Excluding α and β*

For every LDA model created, the minimum word document frequency was two documents, and the maximum word document frequency was 95% of the total documents. These parameter settings are generally standard practice when using LDA models. In addition, The chunk size was set to 100 and the number of passes to 10. Consequently, updating was done 10 times for the book models and 120 times for the chapter models. I chose a high number of passes to account for the small number of documents in the book models but a moderate chunk size so the chapter models would be developed in a reasonable amount of time.

B. α and β Selection Importance

When using the LDA algorithm, there are three critical parameters to consider; the number of topics to generate, α, and β. According to Haaya Naushan [9], α represents the topic density in a document. The higher α is, the more topics coexist in a document. β represents the topic word density. The more words the topic contains, the higher β is. However, this is only true if the topic distribution in the documents is symmetric. If the topic distributions are asymmetric, a higher β results in topics with similar words.

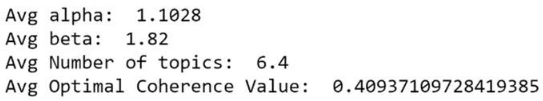
C. α and β Selection Testing; Methodology

I used the LDA model from the Gensim library in Python. I decided to use book documents from the Bible in Basic English translation for my initial testing because I didn't want to use many resources to produce LDA models for the chapter documents. Furthermore, I used the Bible in Basic English translation instead of the King James translation because I assumed the Bible in Basic English would contain more common words, making it easier to interpret the topic word results. I then performed the preprocessing outlined in section III B.

D. α and β Selection Testing; Assuming a Symmetric Model

To determine the best parameter selection for a symmetric model, I created LDA models for every possible combination of the number of topics, α, and β values. The parameters with the highest coherence value were selected as the best model parameters.

For this test, the possible number of topics was two to 10 with a step size of two. The possible α, and β values used were [0.01, 0.1, 0.5, 1.0, 2.0, 3.0]. (These ranges were determined by a few initial tests to see if performing LDA on the Bible would produce practical results). I then computed the optimal parameters 25 times and averaged the values. The average optimal parameters and coherence value from this test can be seen in Figure 5.



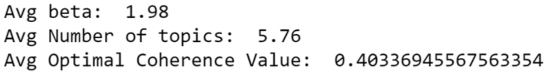
```
Avg alpha:  1.1028
Avg beta:  1.82
Avg Number of topics:  6.4
Avg Optimal Coherence Value:  0.40937109728419385
```

Fig. 5. The optimal parameters and corresponding coherence value for the symmetric model

E. α and β Selection Testing; Assuming an Asymmetric Model

Much like in subsection B, I used Gensim to create LDA models for every possible combination of the number of topics and β values. For the asymmetric model, α is set to auto. Therefore the alpha parameter does not have test values. It should be noted that Gensim allows for two different kinds of asymmetric models, one where the user provides the α values for each document and another where the optimal α values for each document are calculated for the user based on the document contents. I opted for the latter option because I was not sure what α value would be best for each book of the Bible.

For this test, the possible number of topics was two to 10 with a step size of two. The possible β values used were [0.01, 0.1, 0.5, 1.0, 2.0, 3.0], like the symmetric model. I then computed the optimal parameters 25 times and averaged the values. The average optimal parameters and coherence value from this test can be seen in Figure 6.



```
Avg beta:  1.98
Avg Number of topics:  5.76
Avg Optimal Coherence Value:  0.40336945567563354
```

Fig. 6. The optimal parameters and corresponding coherence value for the asymmetric model

F. Evaluating Performance Differences Between Symmetric and Asymmetric Models

According to Figures 5 and 6, neither model outperforms the other based on coherence and the number of topics (If the topic numbers were rounded in the above figures, they would be the same) alone. The β values are also similarly high, which means the topics, in either case, would contain more words from the vocabulary, and in the asymetric case, the words in each topic would be more similar. The implications of that are still unknown.

For my purposes, it makes more sense to select the asymmetric model to improve performance. The multicore LDA model took 37 minutes to calculate the optimal coherence value for one symmetric model. In contrast, it only took 5 minutes to calculate the optimal coherence value for one asymmetric model. Therefore, using an asymmetric model would allow for more parameter values to be tested, thus allowing for a more robust model.

However, before creating and analyzing asymmetric LDA models for the books and chapters of the King James and the Bible in Basic English translations, I wanted to analyze why using an asymmetric model would provide equal or better performance for my use case. I also wanted to investigate whether there would be any drawbacks of using an asymmetric model over a symmetric model.

According to Naushan [9], "This assumption of symmetry would mean that each topic is evenly distributed throughout a document, whereas for an asymmetric distribution (as measured by skewness), certain topics would be favoured over others." Given this definition of an asymmetric distribution, it makes sense why an asymmetric model would be a safer option than the symmetric model, given that the Bible was written in a storytelling format. Therefore, there is no guarantee that all the topics in each book would be evenly distributed.

For example, the book of Psalms has 150 chapters. Each chapter has a different theme and is essentially independent of the other chapters because they are all individual songs or prayers to and for God. One example of the separation of Psalms by theme, produced by The Daily Office of the Catholic Church

TABLE II. PSALMS CHAPTERS SEPARATED BY THEME

Theme	***Verses***
God the Creator.	8, 19, 33, 65, 111, 104, 145, 147.
God the Redeemer.	15, 33, 102, 103, 111, 113, 114, 126, 130, 138.
God the Judge.	1, 7, 11, 46, 50, 62, 75, 76, 82, 90, 96, 97, 98.
God's Glory.	18, 29, 99, 36, 46, 148, 150.
God's Sovereignty.	24, 46, 47, 72, 89, 93, 96, 97, 98, 99, 112, 146, 145.
God's Wisdom.	33, 104, 111, 113, 139, 145, 147.
God's Law.	19, 50, 62, 111, 119, 147. 23, 33, 34, 37, 89, 1 21, 124, 139, 145, 146, 147.
God's Mercy.	23, 32, 57, 61, 62, 63, 73, 77, 85, 86, 100, 103, 118, 130, 145.
The Incarnation.	2, 8, 85, 89, 102, 110, 111, 113, 132.
The Passion.	22, 40, 42, 54, 69, 88, 116, 130.
The Church.	46, 48, 84, 111, 122, 133, 147.
Worship.	5, 26, 43, 63, 65, 66, 67, 84, 96, 100, 102, 116, 122, 138.
Thanksgiving.	30, 65, 67, 92, 98, 100, 111, 103, 107, 116, 13 4, 138, 145, 147, 148, 150.
Prayer.	4, 5, 17, 20, 28, 31, 54, 61, 84, 86, 102, 141, 142.
Trust in God.	27, 31, 57, 62, 63, 71, 73, 77, 91, 118, 121, 123, 124, 125, 143, 146.
God our Refuge.	4, 17, 20, 37, 46, 49, 54, 61, 71, 91, 103, 121, 146.
Divine Guidance.	25, 43, 80, 85, 111, 112.
In Time of Trouble.	3, 11, 12, 13, 18, 20, 30, 40, 46, 49, 57, 62, 63, 80, 85, 86, 90, 107, 118, 144, 146.
Righteousness.	1, 11, 12, 15, 18, 19, 26, 34, 40, 92, 111, 112.
Peace.	29, 46, 76, 85, 98, 100, 124, 125, 126.
The Transitoriness of Life.	39, 49, 90, 102.
The Hope of Immortality.	16, 30, 42, 49, 66, 73, 103, 116, 121, 139, 146.
Morning.	3, 5, 20, 63, 90, 143.
Evening.	4, 13, 16, 17, 31, 77, 91, 121, 134.
Penitential Psalms.	6, 32, 38, 51, 102, 130, 143.
Preparation for Holy Communion.	23, 25, 26, 36, 41, 43, 63, 84, 85, 86, 122, 130, 133, 139.
Thanksgiving after Holy Communion.	8, 15, 18, 19, 27, 29, 30, 34, 100, 103, 110, 11 8, 145, 150.

according to the Anglican Use [10] can be seen in Table II. One LL

can see that there are many significant themes throughout Psalms, and they aren't all necessarily mentioned in other books of the Bible. In the symmetric model, the α parameter was relatively high, meaning that the model assumed that each book of the Bible would contain the same topic distribution when that probably isn't the case.

Now, I needed to know that an asymmetric model would consistently perform equal to or outperform a symmetric model when given various document makeups from different Bible translations.

According to Wallach et al. [11], LDA models with an asymmetric α consistently outperform models with a symmetric α. In their experimentation, they performed topic modeling on various news articles. Findings from their work are referenced in Figure 7. The two models that I'm concerned with are when α and β are symmetric (SS), and when α is asymmetric and β is symmetric (AS). According to their findings, it is generally safer to use an asymmetric α over a symmetric α because it produces minor variation in the topics found within documents. This is presented in Figure a. Figure b shows that as the number of topics increases, an asymmetric distribution slowly divides the original topics. In contrast, a symmetric distribution tends to divide the topics more evenly. Thus, their concluding argument is that it is safer to use asymmetric LDA models because one can overestimate the number of topics to produce without losing as much of the significant topic meanings.

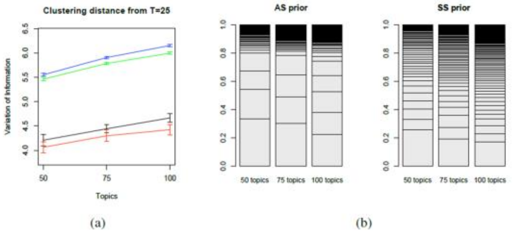


Figure 4: (a) Topic consistency measured by average VI distance from models with $T = 25$. As T incr AS (red) and AA (black) produce $\mathcal{Z}$s that stay significantly closer to those obtained with $T = 25$ th (green) and SS (blue). (b) Assignments of tokens (patent abstracts) allocated to the largest topic in a 25 model, as T increases. For AS, the topic is relatively intact, even at $T = 100$: 80% of tokens assigned topic at $T = 25$ are assigned to seven topics. For SS, the topic has been subdivided across many more to

Fig. 7. The figure from [11]

In addition, Syed et al. [12] performed topic modeling experiments on research articles about fisheries. According to their findings, in most cases, an asymmetric model significantly outperformed a symmetric model in terms of coherence value averages and standard deviations for different topic numbers. In most other cases, the asymetric model performed just as well as the symmetric model. The asymmetric model also performed equal to the symmetric model compared to how humans organized the documents by topic.

After looking at all the information presented above, I determined it was best to use asymmetric models when performing LDA on my data.

G. *Parameter Selection on All Data Using Asymmetric LDA Models*

To determine the best parameter settings for asymmetric models of the King James and the Bible in Basic English translations, I followed the same experimentation method presented in subsection C. For all four models (see section IV), the possible topic range was from two to 66 with a step size of two. This range was selected because I assumed each book would have at most one dominant topic that might be minor in other books. The possible β range was 0.01 to 3.0, with a step size of 0.05. This range was selected because β averaged comfortably in this range in the testing in subsection E. α was set to auto. The optimal parameter settings (for the number of topics and β) and corresponding coherence value for all four models is shown in Table III.

I then wanted to find the average optimal parameter settings for each model. Using the parameter settings in Table III as initial estimates, I again searched for the optimal coherence value, but in a smaller range around those parameters for each model. The new parameter ranges tested for each model are shown in Table IV. The possible topic range is -5/+5 of the numbers presented in Table III, and the possible β ranges between -0.9/+1.3 of the numbers presented in Table III. I then calculated the optimal coherence value and corresponding parameter settings for each model 50 times. Finally, I averaged the optimal parameter values, which became my final model parameters (the average number of topics are rounded to the nearest whole number in the final models). These can be seen in Table V.

TABLE III. THE INITIAL OPTIMAL PARAMETER SELECTIONS AND CORRESPONDING COHERANCE VALUES FOR EACH ASYMETRIC MODEL

Model	***Coherence Value***	***Number of Topics***	***β***
BBE Books	0.42455	8	1.81
BBE Chapters	0.8106	28	2.91
KJV Books	0.4513	8	1.81
KJV Chapters	0.8060	22	2.61

TABLE IV. THE SEARCH PARAMETERS FOR DETERMINING THE OPTIMAL PARAMETERS OF EACH ASYMETRIC MODEL, BASED ON THE PARAMETERS IN TABLE II

Model	***Topics Range***	***β Range***
BBE Books	3-13, step size: 2	0.91-3.11, step size: 0.3
BBE Chapters	23-33, step size: 2	2.01-4.21, step size: 0.3
KJV Books	3-13, step size: 2	0.91-3.11, step size: 0.3
KJV Chapters	17-27, step size: 2	1.71-3.91, step size: 0.3

TABLE V. THE AVERAGE OPTIMAL PARAMETER VALUES AND CORRESPONDING AVERAGE COHERENCE VALUES OVER 50 EXECUTIONS TO DETERMINE THE OPTIMAL PARAMETER VALUES

Model	***Average Coherence Value***	***Average Number of Topics***	***Average β***
BBE Books	0.6266	7.84	2.28
BBE Chapters	0.8197	26.92	3.47
KJV Books	0.6255	7.68	2.04
KJV Chapters	0.8041	18.8	3.19

VI. EXPERIMENTATION AND RESULTS

A. *Interpretation Of Parameter Selection Results*

As one can see from Table V, there isn't much performance difference in the β or coherence values when comparing the models by document makeup. The notable difference is that the King James chapter model produced roughly 19 topics, whereas the Bible in Basic English model produced 27 topics. This difference is discussed in later sections.

When comparing the coherence values, it is self-evident that using the chapter documents significantly increases the coherence value. This means document makeup is essential for model performance. Having smaller documents allows for more specific word correlations because fewer words could form weak correlations with each other. Having more documents allows for stronger word correlations because there are more opportunities to find the exact word correlation in multiple documents. Both of these principles increase the coherence value.

B. *Topic Visualization Using Optimal Parameters*

Figure 8 shows four LDA model visualizations using the parameters discussed in section V. Each circle in the figures represents a topic. The circle size represents the marginal topic distribution, and the distance between the topics represents their semantic similarities. It should be noted that while there is only one model for each translation and document combination, multiple LDA models were created under the previously mentioned parameters, and the majority of them produced similar results. The corresponding coherence values for each model are shown in Table VI.

TABLE VI. THE CORRESPONDING COHERENCE VALUES FOR THE MODELS IN FIGURE 8

Model	***Coherence Value***
BBE Books	0.3618
BBE Chapters	0.7142
KJV Books	0.3320
KJV Chapters	0.6539

All four models seem to produce visually similar results. There are approximately one or two dominant topics. The rest of the topics have much semantic overlap, and they are less prevalent in the marginal topic distribution. This may be considered a negative aspect of these models. However, more testing would have to be done to see if that is the case. The smaller topics may still be very prevalent in the Bible, but only in one book or a subset of chapters. In addition, many of the most frequent words in all four models are very similar. This is positive because it shows that the models are picking up on the general topics in the Bible and corroborating each other.

It is also interesting to note that the average β values from Table V reflect the models' visualizations. As mentioned previously, when using an asymmetric model, a higher β means that the topics consist of similar words. According to Figure 8, many of the models' topics overlap, meaning they are semantically similar. Since the chapter models have more topics that overlap, compared to the book models, it makes sense that they would have higher β values. Similarly, it makes sense that the book models would have close β values because they have roughly the same number of overlapping topics.

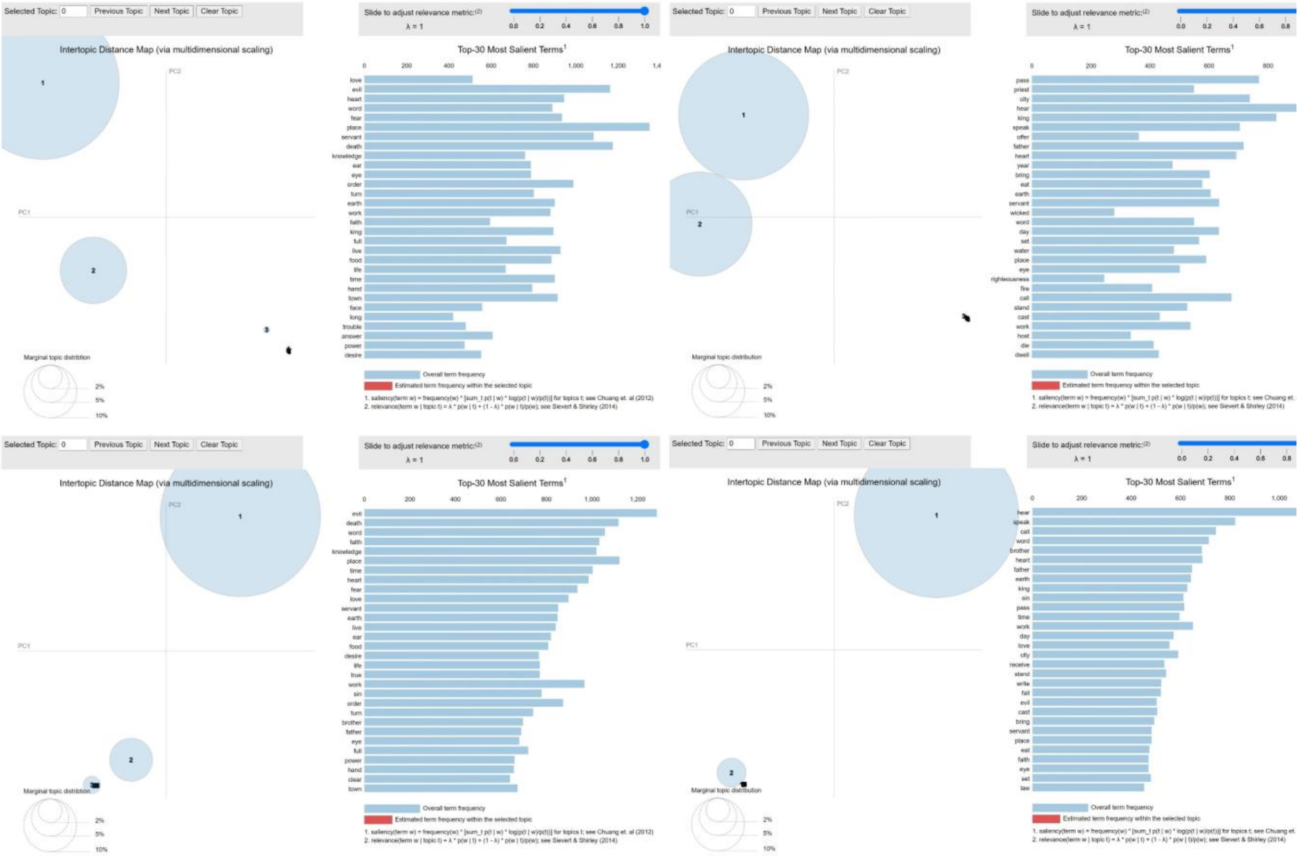

Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Intertopic Distance Map (via multidimensional scaling)
PC1
PC2
1
2
3
Marginal topic distribution
2%
5%
10%
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Top-30 Most Salient Terms1
0
200
400
600
800
1,000
1,200
1,4
love
evil
heart
word
fear
place
servant
death
knowledge
ear
eye
order
turn
earth
work
faith
king
full
live
food
life
time
hand
town
face
long
trouble
answer
power
desire
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)
Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Intertopic Distance Map (via multidimensional scaling)
PC1
PC2
1
2
Marginal topic distribution
2%
5%
10%
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
Top-30 Most Salient Terms1
0
200
400
600
800
pass
priest
city
hear
king
speak
offer
father
heart
year
bring
eat
earth
servant
wicked
word
day
set
water
place
eye
righteousness
fire
call
stand
cast
work
host
die
dwell
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et.
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)
Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Intertopic Distance Map (via multidimensional scaling)
PC1
PC2
1
2
Marginal topic distribution
2%
5%
10%
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Top-30 Most Salient Terms1
0
200
400
600
800
1,000
1,200
evil
death
word
faith
knowledge
place
time
heart
fear
love
servant
earth
live
ear
food
desire
life
true
work
sin
order
turn
brother
father
eye
full
power
hand
clear
town
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)
Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Intertopic Distance Map (via multidimensional scaling)
PC1
PC2
1
2
Marginal topic distribution
2%
5%
10%
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
Top-30 Most Salient Terms1
0
200
400
600
800
1,000
hear
speak
call
word
brother
heart
father
earth
king
sin
pass
time
work
day
love
city
receive
stand
write
fall
evil
cast
bring
servant
place
eat
faith
eye
set
law
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et.
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 8. The visualization of the four LDA models using the parameters discussed in section V. The figure in the top left is the Bible in Basic English book model. To the right of that is the King James book model. The figure in the bottom left is the Bible in Basic English chapter model. To the right of that is the King James chapter model.

Figures 9-12 show the most frequent words for the first two topics in each book model. The first figure for each topic displays its least exclusive words, and the second figure displays its most exclusive words. As one can see, the first and largest topic in both translations seems to focus on generalities in the Bible using words like faith, love, righteousness, wisdom, hope, and scripture. The second topic in both translations seems to focus on the ideas of worshiping God and community, especially concerning the Old Testament; using words like King, priest, offer, eat, brother, daughter, sin offering, burnt offering, unclean, silver, and gold. Many of these ideas are related to the offerings that people would give to God during the Old Testament. Such as the burnt offering of various animals. The offerings would be burned in the tabernacle or tent. In addition, the Levites would be able to eat the burnt offerings that people brought. Silver and gold are also mentioned frequently because silver and gold items were frequently considered offerings to God and were used in many places throughout the tabernacle.

These topics demonstrate how the LDA models function practically, and I am pleased with the results. Even though there are many overlapping topics, the two topics mentioned in the previous paragraph are fundamental aspects of the Bible. It is also good to see that the models agree on the correlated words for those topics. Even though I did not present the most frequent words for the largest two topics in the chapter models here, I did observe them, and they seem to agree with the first two topics in the book models.

Upon a more in-depth observation, one can see that the Bible in Basic English model seemed to produce better topic exclusivity results for topic one, and the King James model produced better topic exclusivity results for topic two. In this case, better exclusivity is defined as the least exclusive words in the topic being more likely to belong to only that topic. Visually this means the translation's leftmost figure had the smallest overall term frequency relative to the frequency within the topic.

When comparing these results to the visualization in Figure 8, this intuitively makes sense. In either case, the topic in question has the largest marginal topic distribution in its respective translation. This means that the topic appears more in that translation over the other. Therefore, the difference in term frequency is occurring for each topic.

I suspect that the Bible in Basic English translation performed better in topic one because the stop words I used were some of the most frequent words in the King James translation of the Bible. This means that the King James models lost many of the words that would appear

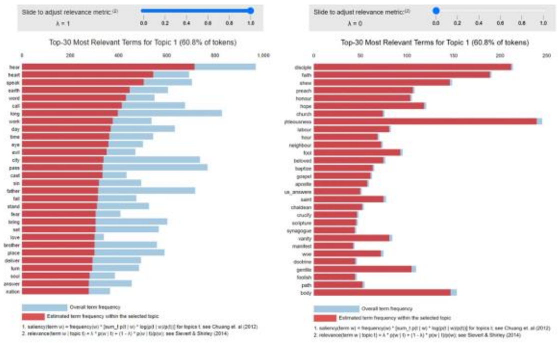


Fig. 9. The least and most exclusive words for the first topic in the King James book model

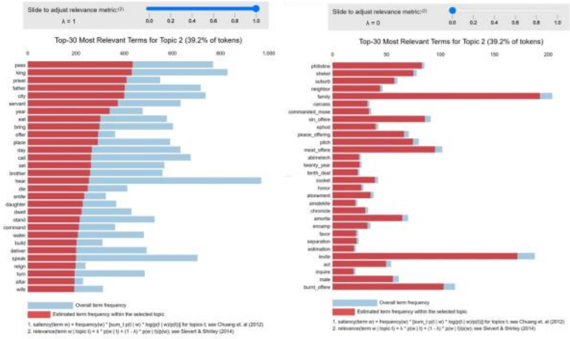


Fig. 10. The least and most exclusive words for the second topic in the King James book model

frequently throughout the documents, thus resulting in lower word frequencies. Regarding topic two, the King James model contains more general family-oriented words than topic two in the Bible in Basic English model. These words appear more frequently in the Bible in general, which might be why the King James model has a larger marginal topic distribution for topic two. The King James model also includes fewer words related to the idea of worshiping God relative to the Bible in Basic English model. Those words could be less frequent than the other topic two words in the King James model. These visual observations don't necessarily show better performance of one model over the other. However, they demonstrate the importance of human interpretation and analyzing use cases when using LDA models. Either model could be useful when looking at these two topics, depending on how general the topics need to be.

Even though a visual inspection of the models seems optimistic, there is still much uncertainty in the models that I cannot analyze in this paper simply because of the number of topics in each model. However, this uncertainty is potentially resolved or mitigated when looking at the models' coherence values. Interestingly, the coherence values are approximately 30 percent below the corresponding book model averages and 10 to 15 percent below the corresponding chapter model averages in Table V. This could mean that I needed more data to arrive at the appropriate parameter values and optimal coherence value. On the other hand, it could mean that averaging the coherence value wasn't the best way to determine the most common optimal parameters.

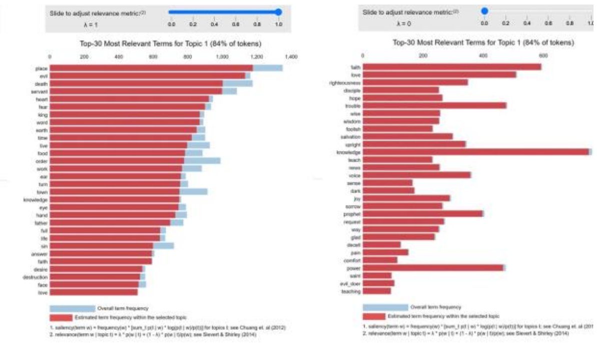


Fig. 11. The least and most exclusive words for the first topic in the Bible in Basic English book model

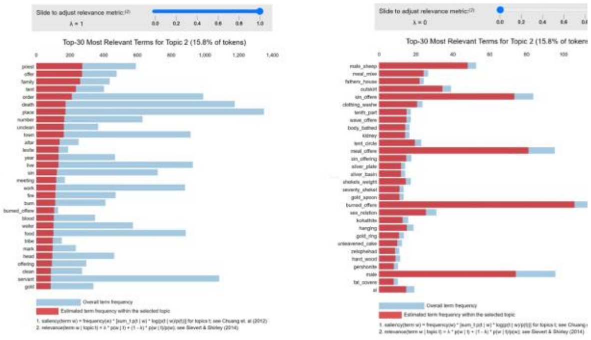


Fig. 12. The least and most exclusive words for the second topic in the Bible in Basic English book model

In either case, based on the coherence values, the Bible in Basic English models produced better results than the King James models. This can also be seen in Figure 8 because the Bible in Basic English models have one or two more topics than the King James models that are larger and do not overlap with many other topics. Therefore, while the initial inspection of Figure 8 didn't definitively show that one set of models performed better over another when referencing Figure 8 in conjunction with the coherence values, one can really understand the performance differences.

The difference in coherence values of the translations could be attributed to the fact that the Bible in Basic English chapter model has more topics than the corresponding King James model. According to the findings from [12], the coherence value tends to increase as the number of topics increases. That likely occurred with the chapter models because the book models had the same number of topics and similar coherence values. However, I cannot be positive because the book and chapter models have more overlapping topics than those in [12], which might affect the coherence values.

VII. Conclusions and Future Work

A. Conclusions

While I think all the models presented in the previous sections of this paper prove that LDA can be performed on the whole Bible with good results, there is room for improvement. Notably, having the document makeup consistent of Bible chapters produced much better coherence values than the document makeup consisting of books. However, it is also vital to notice that while the models might not be optimal, they seem similar when comparing them by document makeup.

This paper also effectively highlights the importance of preprocessing documents when using LDA models, especially in a story-based, semantically classical text. Stop words need to be investigated when working with texts that use more traditional language. It also shows the need to determine if the uncommon words in documents are in the model vocabulary. If many of the uncommon words in the documents are relevant, natural language processing techniques may be less practical.

In addition, the similarities and differences between Bible translations can be seen in multiple aspects of this paper. From a classification standpoint, the translations of the Bible selected for the LDA models appear to be very different, but the LDA model topics were very similar. Thus, potentially topic modeling only needs to be performed on one translation of the Bible to sufficiently describe any translation of the Bible, despite their semantic differences. However, the models did differ slightly concerning how general (or widespread) a topic was. Therefore it is always essential to determine the use case when selecting documents for an LDA model. It is also important to note that it was only feasible to compare two model topics in this paper. Thus these observations, while optimistic, are not thoroughly supported.

My work also reinforces the idea that asymmetric LDA models can produce equal if not better performance than symmetric models. Not only do asymmetric models keep the user from having to determine α, but they also provide more leeway when determining the optimal number of topics. In addition, asymmetric models tend to provide better performance from a hardware-based standpoint because they execute faster than symmetric models. However, it is evident that β plays a considerable role in asymmetric model performance because larger β values can lead to overlapping topics.

B. *Future Work*

In the future, it might be interesting to test models using different biblical stop word lists to see which words produce better performance without taking away essential words from topics. It might also be interesting to experiment with the chunk size and passes parameters of the LDA models to see how that affects results since those weren't investigated in this paper.

In addition, since many of the topics in the models overlap semantically, determining optimal parameters using different methods might be worth exploring. For example, instead of only looking at the optimal coherence value, one could observe the trends in the coherence value as the parameters change and determine the optimal value based on those trends.

It may also be useful to determine the most common parameters of the optimal coherence values instead of averaging the parameter values. While the average should eventually approach those most common values, it is impossible to know how much time that would require.

It may also be beneficial to experiment with smaller β values or a smaller number of topics to see how that affects topic overlap and the optimal coherence value. Even if a model has a lower coherence value, it may be better than the current models because the topics might vary more in terms of the words they contain.

References

[1] The T. Mumford, "Literary mysteries: The best-selling books of All time," *MPR News*, 21-Jul-2015. [Online]. Available: https://www.mprnews.org/story/2015/07/21/thread-books-bcst-best-selling-books. [Accessed: 05-Nov-2021].

[2] "Religion in America: U.S. religious data, Demographics and statistics," *Pew Research Center's Religion & Public Life Project*, 09-Sep-2020. [Online]. Available: https://www.pewforum.org/religious-landscape-study/frequency-of-reading-scripture/. [Accessed: 05-Nov-2021].

[3] V. K. Garbhapu, "A comparative analysis of latent semantic analysis and latent Dirichlet allocation topic modeling methods using Bible Data," *Indian Journal of Science and Technology*, vol. 13, no. 44, pp. 4474–4482, 2020.

[4] W. Hu, "Unsupervised learning of two Bible books: Proverbs and psalms," *Sociology Mind*, vol. 02, no. 03, pp. 325–334, Jul. 2012.

[5] O. R. Hartono, "Bible corpus," *Kaggle*, 16-Jun-2017. [Online]. Available: https://www.kaggle.com/oswinrh/bible. [Accessed: 05-Nov-2021].

[6] B. Haugard , "Biblical Names and their Meanings.".

[7] W. J. Chamberlin, *Catalogue of english bible translations: A classified bibliography of versions and editions including books, parts, and Old and new testament apocrypha and Apogryphical Books*. Greenwood, 1991.

[8] "Popular bible words," *POPULAR BIBLE WORDS*. [Online]. Available: https://www.kingjamesbibleonline.org/Popular-Bible-Words.php. [Accessed: 05-Nov-2021].

[9] H. Naushan, "Topic modeling with Latent Dirichlet allocation," *Medium*, 02-Dec-2020. [Online]. Available: https://towardsdatascience.com/topic-modeling-with-latent-dirichlet-allocation-e7ff75290f8. [Accessed: 05-Nov-2021].

[10] "A table of psalms by theme," *The Daily Office of the Catholic Church according to the Anglican Use*. [Online]. Available: http://www.bookofhours.org/psalms/tool_themes.htm. [Accessed: 05-Nov-2021].

[11] H. Wallach, D. Mimno, and A. McCallum. 2009. Rethinking LDA: why priors matter. In Proceedings of NIPS, pages 1973–1981.

[12] S. Syed and M. Spruit, "Exploring symmetrical and asymmetrical Dirichlet priors for latent Dirichlet allocation," *International Journal of Semantic Computing*, vol. 12, no. 03, pp. 399–423, 201