Sacred Text Similarities: A NLP and Machine Learning Approach

Warren Geither
Virginia Tech
Blacksburg, VA
wgeither@vt.edu

Abstract

Although a stark contrast at first, technology and religion share many of the same characteristics. Both give people the power to create, come together, and better the world. On the other side, they give people the power to destroy, segregate, and push society in a degenerative direction. Thus, to keep us as a species moving forward, it is essential that we understand both of these forces. Lucky for us, we can use one of these to help us explore the other. In this project, we will use Natural Language Processing and Machine Learning techniques to analyze 5 sacred texts across different religions (Bible, Book of Mormon, Gospel of Buddah, Koran, Meditations by Marcus Aurelius). We find that while all texts share similar sentiments, some are found to be more similar than others as seen through consine similarity and classification tasks.

1 Introduction

Throughout history we have seen the power of religion influence almost every dimension of life from one's personal life to the greater socioeconomic/cultural climate of the entire world. We've seen awful atrocities such as the crusades and terrorist attacks taken in the name of a higher power to beautiful Sunday services that bring communities together each week.

All of the tensions that come between various religions can largely be attributed to focusing on their differences. However, what if they have more in common than meets the eye? Taking the objective stance of our lord and savior "Science", we can take an unbiased look at the similarities and differences between religions/world views. Specifically we can use natural language processing and machine learning to investigate the supposed dichotomies.

Others have used these techniques to explore similar texts. [1][2][3] This paper will attempt to recreate some of their methods while also attempting some novel approaches such as VADER sentiment analysis, sentence classification, and text generation.

1.1 Data

The dataset being used is a Kaggle dataset containing text files of the 5 sacred texts mentioned above [4]. These e-books were taken originally sourced from the Project Gutenberg website, an online book archive.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

1.2 Data pre-processing

As with every NLP analysis, we must perform some data cleaning, specifically tokenizatoin and removing stopwords.

Different analyses required different tokenization of the texts, so the text was parsed by sentences, words, and characters. In addition, a general list of stopwords from the nltk package was used as a filter with some custom ones that are specific to the old language used in these texts such as (thou, shall, shalt, etc.).

2 Methods and results

2.1 Exploratory data analysis

Once we have preformed our data pre-processing, EDA was conducted to understand the data further.

Word clouds are often the first tool in the arsenal of someone working with text data. They make it easy to visualize and understand the distribution of words within a given text. The more frequent the word in the document, the larger it will appear in the cloud. Figure 1 and 2 show two examples.

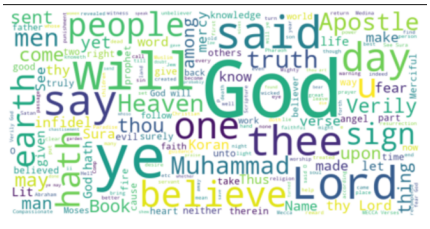

Figure 1: Wordcloud from the Koran

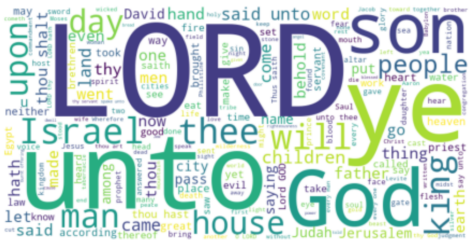

Figure 2: Wordcloud from the Bible

2.2 VADER sentiment analysis

VADER stands for "Valence Aware Dictionary for Sentiment Reasoning." It is a model that uses the emotional intensity of words in a sentence to produce sentiment scores. This model is specifically attuned to social media text [5].Although social media is far removed from these sacred scriptures, it should provide a good baseline for sentiment comparison between these texts.

The scores given are the positive, negative, and neutral value of a sentence or word along with an overall comprehensive score that encompasses all of the values. Figure 3 shows our results.

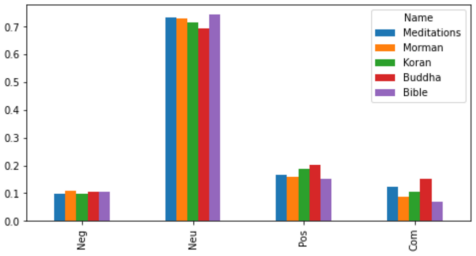

Figure 3: Results from sentiment analysis

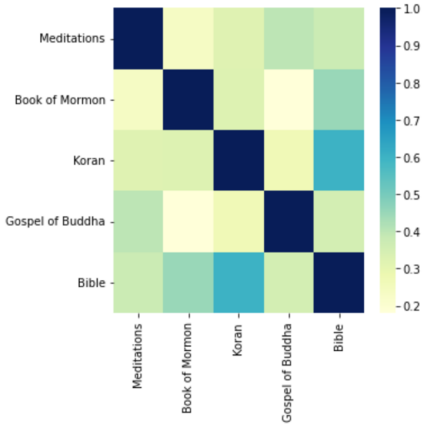

Figure 4: Cosine similarity matrix

2.3 Cosine similarity matrix

To further assess how similar these texts are, the cosine similarity was found between each pairing of texts. Cosine similarity is typically the distance metric most used with text data since it is less affected by repeated words or text length than other distance metrics like euclidean distance. Figure 4 shows a heatmap for all the different pairings of texts.

2.4 Word2Vec

The Word2Vec algorithm uses a nearual net to learn the association between words. It was used to determine the most similar words in terms of Euclidean distance to "god". Table 1 has the most related words in each book.

Table 1: Most similar words to "god"

Category		
Text	Word	Distance
Bible	redeemer	0.791
Meditations	men	0.999
Mormons	statutes	0.924
Koran	trust	0.961
Buddah	family	0.999

2.5 Sentence classification

Four different classification models: Naive Bayes, KNN, SVM, and Random Forest were trained on the sentences of each of the texts to predict the label. The Naive Bayes confusion matrix is shown in Figure 6, however all models performed similarly. The model performed with an accuracy of 0.87 and had a macro precision and recall of 0.86 and 0.8 respectively. Mapping = {'Bible': 0, 'Book of Mormon': 1, 'Gospel of Buddha':2, 'Koran':3, 'Meditations':4}

Figure 5 shows some class imbalance, but this did not seem to have a large adverse affect on our models performance.

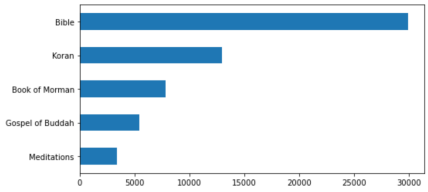

Figure 5: Number of sentences per book

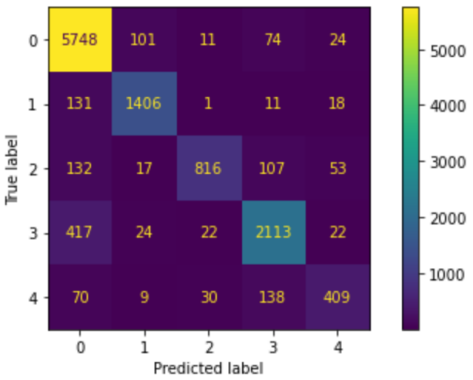

Figure 6: Heatmap of naive bayes confusion matrix

2.6 Text generation with LSTM

Long-Short Term Memory (LSTM) Neural Networks are a type of recurrent neural network which has the ability to take into consideration past data during the training phase. This results in much better performance than RNN and fixes the common vanishing or exploding gradient problem.

Ideally, this segment was intended to train a Long-Short Term Memory (LSTM) Neural Network on all 5 texts and see what kind of texts are generated when a sequence like "God is..." is presented to the model. However, due to the length of time it would take to train, only 1 text was used as a proof of concept.

The training took 2 days for 100 epochs and in the results there was a common phenomena of neural text degeneration.[6] This came in the form of repeating characters "a" and "t" when given a root sequence. Further work is needed to overcome this hurdle.

2.7 LDA topic analysis

Latent Dirchlet Allocation(LDA) is a popular technique to preform topic analysis on a corpus of texts. It assumes that a document is made of a mixture of topics and that topics are made up by a mixture of words. Named for using a Dirchlet distribution which is a probability distribution whose realizations are probability distributions. We can view the document as a probability distribution over the topics and the topics as a probability distribution made up of words. Figure 7 shows a visualization of the LDA procedure completed on the texts. The topics are shown as the circles. The words to the right are the top 30 terms that make up the red circle.

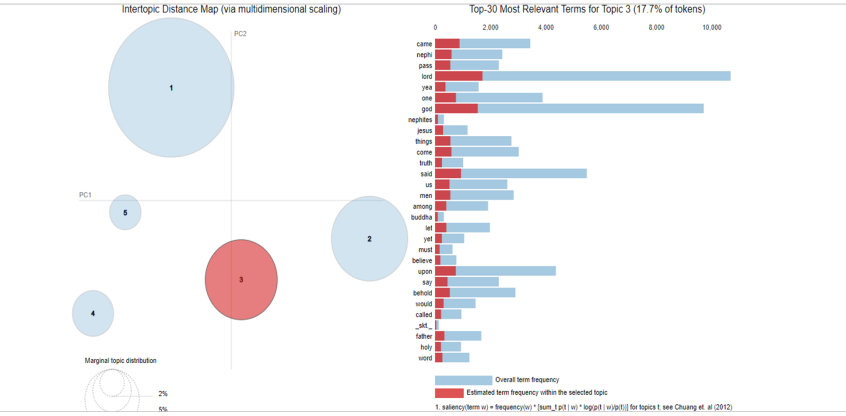

Figure 7: LDA topic analysis

3 Discussion

Wordclouds From visual inspection of the word clouds in Figure 1 and 2, we can see that the Koran and Bible seem to have some words that are used in similar frequencies such as "Lord" and "God". All other scriptures are relatively easy to distinguish although the wordclouds are not displayed her due to space requirements. However, this similarity of words between the Koran and Bible is interesting given the tension between the groups.

Cosine similarity What we see in the wordclouds is further supported by our Cosine Similarity matrix. We find that the Bible and Koran have the highest cosine similarity. Another interesting observation is that The Book of Mormon and The Gospel of Buddha have the least in common. This again highlights that some text are more similar than others in terms of the language they use.

Sentiment analysis Through our sentiment analysis in Figure 3, we can see that all of the texts share similar sentiments across the board with neutral being the overarching category. This shows us that on average, not one text is using more negative or positive language than another.

Word2Vec Its easy to see how these terms in Table 1 can all be related to "god" in their own context given the text it came from. If we were to further this analysis I am sure we would find similar word associations amongst the texts.

Sentence classification From the confusion matrix, accuracy, precision, and recall, we can see that our model worked relatively well predicting the labels from the sentences. Our theme continues as

we see the Koran being mislabeled as the Bible having the highest misclassification rate amongst the texts. Overall, our models more than adequate classification performance clearly indicates that it is picking up some differences between the texts.

Text generation Ideally a neural net trained on all of the world's religious texts would give birth to a new unifying view of the universe taking into account all of the similarities and differences. This project is at least a small step toward this.

LDA topic analysis We notice the topic displayed includes words from each of the text: "god" from the Koran, "jesus" from the Bible, "men" from Meditations, "nephi" from "The Book of Mormon", and "buddha" from the Gospel of Buddha. Since the analysis did not silo each text into its own topic, this may suggest that these texts share similar topics. Further analysis of the rest of the topics could give more evidence to this.

4 Conclusion

This paper sought to objectively analzye the similarities and differences between these sacred texts. The wordclouds, cosine similarity matrix, sentiment analysis, and topic analysis show us some similarities between these texts, especially between the Koran and the Bible. While the sentence classification models show that there are indeed some differences between them. Greater awareness of the similarities and differences will foster understanding and unity which will be of benefit to the whole human race.

5 Further research

The analyses here are by no means comprehensive. Successfully debugging the text generation model or perhaps using a Transformer model such as BERT instead of LSTM would produce interesting results to analyze. And I am sure we will see further interest in this area as advancements are made past NLP to NLU (Natural Language Understanding).

Broader Impact

Understanding the similarities and differences in religious texts will prevent war, bring about world peace, and usher a new age of peace & love for mankind.

In terms of bias, only 5 sacred texts are being used opposed to the thousands that are possible. And even within the ones used, there are many different translations that are not accounted for.

Contributions

Warren Geither preformed all analyses.

References

[1] Sah, Preeti. (2019). What do Asian Religions Have in Common? An Unsupervised Text Analytics Exploration.
[2] Peuriekeu, Y.M., Noyum, V.D., Feudjio, C., Goktug, A., & Fokoue, E.P. (2021). A Text Mining Discovery of Similarities and Dissimilarities Among Sacred Scriptures.
[3] Varghese, Nisha & Punithavalli, M. (2020). Lexical And Semantic Analysis Of Sacred Texts Using Machine Learning And Natural Language Processing. International Journal of Scientific & Technology Research. VOLUME 8. 3133-3140.
[4] Rob Harrand. (2020, January). Religious and philosophical texts, Version 2. Retrieved March 1, 2022 from https://www.kaggle.com/datasets/tentotheminus9/religious-and-philosophical-texts.
[5] Hutto, C.J. Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.
[6] Holtzman, Ari, Buys, Jan, Forbes, Maxwell, Choi, Yejin (2019). The Curious Case of Neural Text Degeneration