Daniel McArthur
Natural Language Processing Final Project

Cross-Linguistic Sentiment in the Bible

Abstract

While much research in the field of Natural Language Processing (NLP) has been carried out in English, there is still a wide range of research questions about cross-linguistic NLP that have been left unanswered. For one, what can comparing a text with multiple languages reveal about the culture of the speakers of those languages? The tool for comparison chosen for this project was sentiment analysis, as this is a potentially overlooked area when it comes to translation studies. Thus, the guiding research question became: what can sentiment analysis on literature in translation reveal about the culture of the speakers of those languages? The text chosen was the Bible, primarily due to the amount of translations that are publicly-available. The languages and models used were likewise based on available models and datasets, and the final analysis includes English, Spanish, Italian, German, French, Dutch, Turkish, Arabic, Japanese, Persian, and Indonesian. One model was used for the European languages, while the non-European languages required separate models, leading to potentially inconsistent and limited results. Still, the resulting calculations give insight into the possibilities for this area of research. The results indicated that, despite some largely variable data, all models tended to categorize the same sections of the bibles with the same overall sentiment. Certain books and languages ended up having a large amount of variation, but it is not clear why that is the case. At a deeper level, the language used in high- and low-sentiment verses was generally consistent across all languages, while each individual language had its own particular way in which positive or negative sentiment was shown (i.e. which words tended to have a high or low sentiment). Expanding this to the larger field of NLP demonstrates areas in which cross-linguistic NLP is lacking, and why further research in that area is important.

Introduction

Comparing features of different languages is a powerful tool in for researching the culture of the people who speak that language. For instance, the field of historical linguistics heavily relies on comparison in order to understand proto languages and proto cultures. With this in mind, comparative linguistics that incorporates language sentiment may add another dimension to both linguistics and NLP. Specifically, it could indicate what words or cultural values are associated positively or negatively within and across various languages. In order to get at this question, I decided to look at cross-linguistic sentiment analysis in one text. Comparing how one narrative gets retold across languages could indicate numerous points of variation or similarity in different cultures. The bible was chosen for this analysis not only for its touching on cultural values, but because of the wealth of available parallel texts in the public domain.

Cross-linguistic sentiment analysis also has broader implications in the field of natural language processing. The majority of work in the field deals with the English language, which, while helpful, has come at the cost of overlooking the importance of other languages. Performing this research can highlight the importance and value in cross-linguistic NLP, as well as highlight areas where it is lacking.

Methodology/Dataset

The datasets were taken from biblesupersearch.com, a website that hosts dozens of bible translations in a variety of formats. The languages and editions were limited to what was on the website, but I downloaded versions of the bible in English (King James), Spanish, Italian, German, French, Dutch, Turkish, Arabic, Japanese, Persian, and Indonesian. Given the length of the texts, as well as the number of translations, performing a deep analysis into the accuracy of the source texts was not feasible. However, aside from a few missing/inconsistent datapoints, the datasets looked reliable.

I initially performed an EDA on the English-language dataset in order to gain a better sense of the data I was working with. After making sure the data was cleaned and organized well, I moved onto performing sentiment analysis on the data. I initially chose to use the nlptown/bert-base-multilingual-uncased-sentiment model for analyzing the Dutch, English, French, German, Spanish, and Italian datasets. Despite potential pitfalls, the one model was used to analyze all six languages for a variety of reasons. For one, it was much simpler and easier to use one model that was compatible with six languages. Secondly, any errors or drawbacks from the model would likely be reflected across all six languages (i.e. any major differences in results would not be due to a difference in the model). Primarily, though, I was limited by the models that were available. To try and include a wider range of languages, I included an analysis of Turkish, Arabic, Indonesian, Japanese, and Persian in order to factor in non-European languages. These specific languages were again chosen based on available datasets and models. The models used were CAMeL-Lab/bert-base-arabic-camelbert-da-sentiment for Arabic, christian-phu/bert-finetuned-japanese-sentiment for Japanese, savasy/bert-base-turkish-sentiment-cased for Turkish, and ayameRushia/bert-base-indonesian-1.5G-sentiment-analysis-smsa for Indonesian.

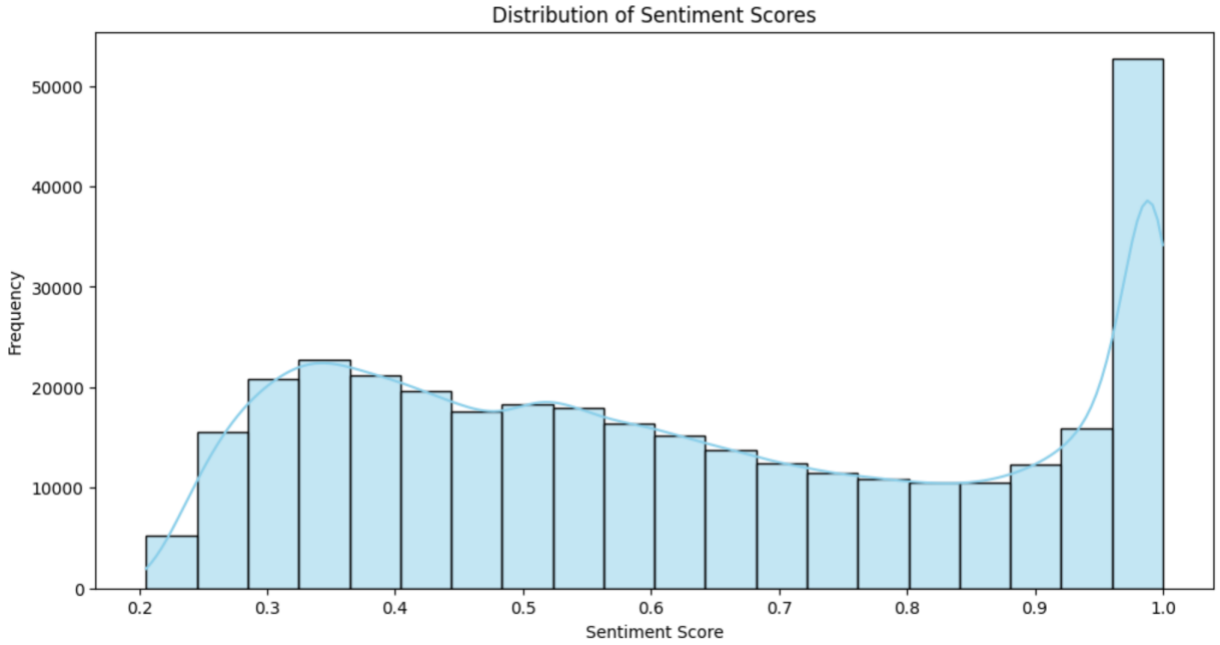
The largest drawback to the models I chose was in their training data, which was mostly on customer reviews. Working with historic and antiquated language meant that there was a lot of different language data that the model would be less confident about. To counteract this, fine-tuning the models would be extremely helpful, but I was unable to find any dataset containing historic data labeled for sentiment in *all languages* being analyzed. Nor was I able to label training data myself, as I am not fluent in all of the present languages. Thus, the analysis might just give a "best guess" at what a more fine-tuned model might predict.

The analyses were done at the verse-level, which was largely done because of the nature of the dataset, which broke down entries into verses. This was not only beneficial in equalizing the length of inputs, but also because of the length of text that the models were trained and fine-tuned on. However, analyzing the verses in isolation can factor out relevant context that might alter the sentiment of an individual verse. Once all of the data was analyzed for sentiment, I ran through a number of different data visualizations to display trends within and across books (books with highest and lowest sentiment, variance across languages, common positive and negative words, etc.). Much of the decisions on relevant visuals and data points throughout the analysis were made on-the-fly, as this project was an immense learning process for me.

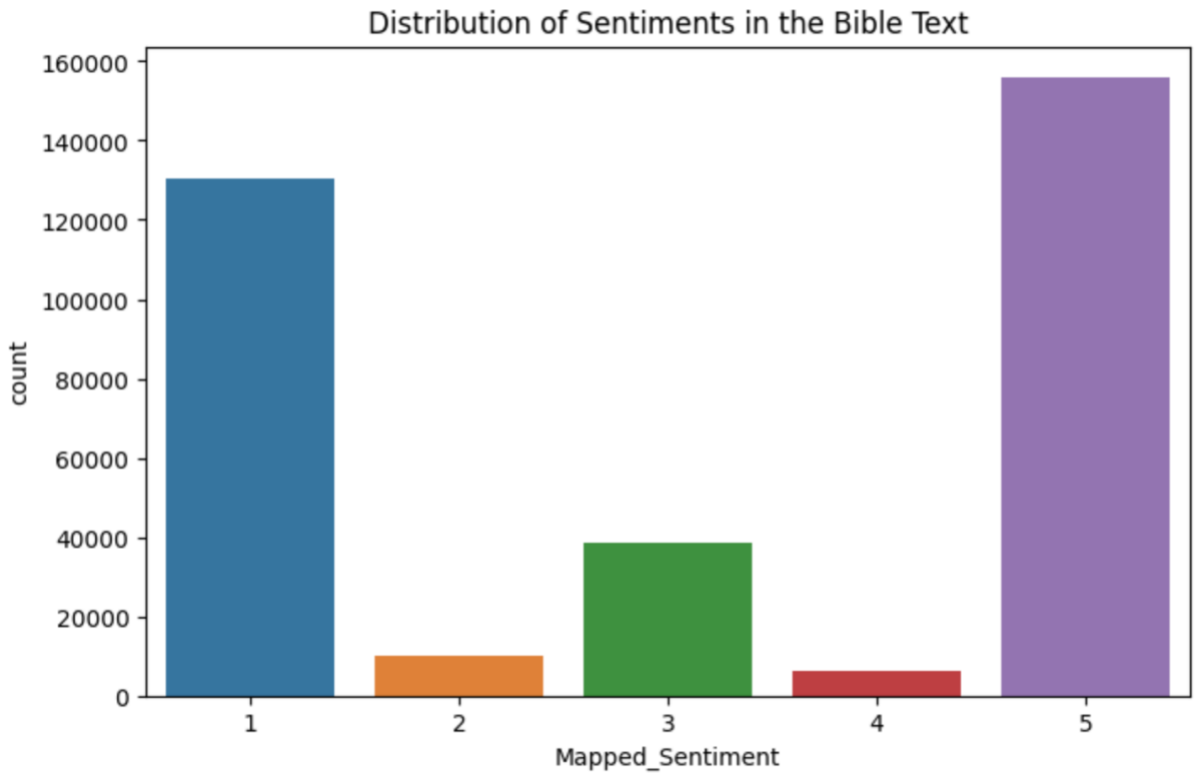
Results

Before analyzing the results of the sentiment analysis, it's important to see how accurate the models themselves are. The multilingual model had an overall accuracy of 45.21% (based on the Sentiment Score from the model, which appeared an evaluation metric). The other models had a comparably higher score, with 77% for Arabic, 93% for Indonesian, 83% for Persian, 83%

for Turkish, and 80% for Japanese. Given that all of these models were also trained on modern language, though, the metrics do not seem accurate. Furthermore, we can see the distribution of sentiment scores per verse across all languages in order to gauge how the models performed.



The measures for sentiment in each model used either a scale of 1-5 (1 being low and 5 being high); ‘Positive’ and ‘Negative’; ‘Positive’, ‘Negative’, and ‘Neutral’; or ‘Happy’ and ‘Sad’. The multilingual model tended to categorize text as either 1 or 5, with relatively few in 2, 3, or 4. The other models had a more even distribution, but tended to skew negative, with Indonesian and Arabic especially leaning negative. Japanese was the only non-European language model that skewed positive. Mapping all models to a scale of 1-5 yielded the following distribution:



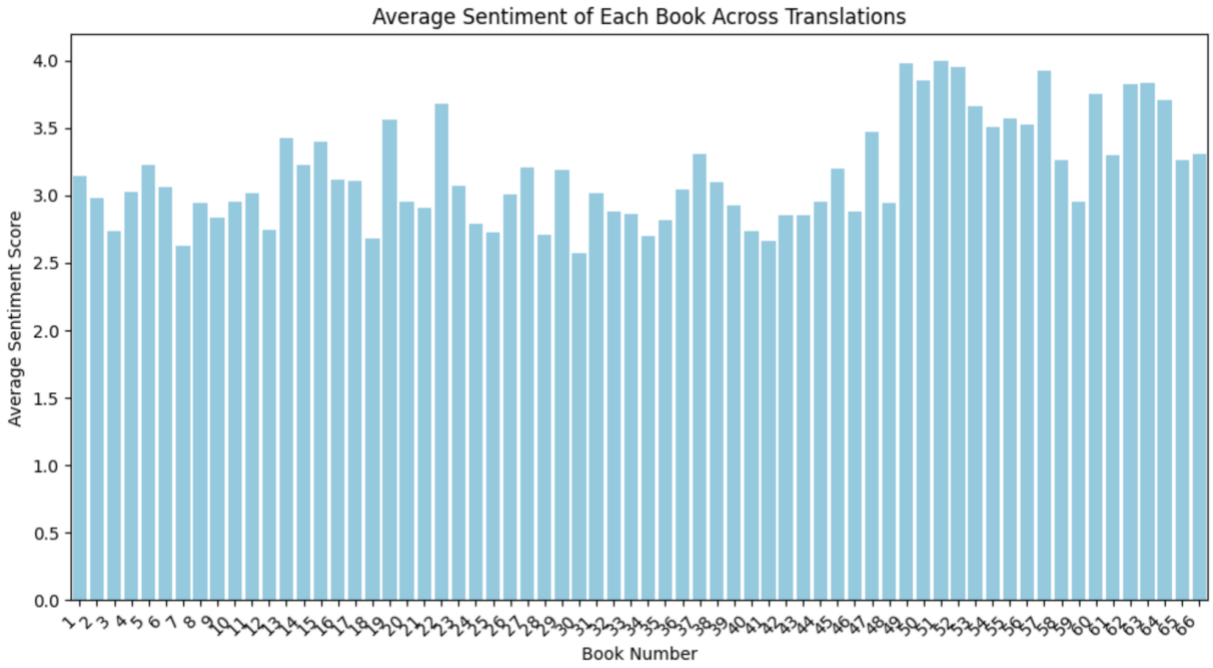
The average sentiments of all verses per language were:

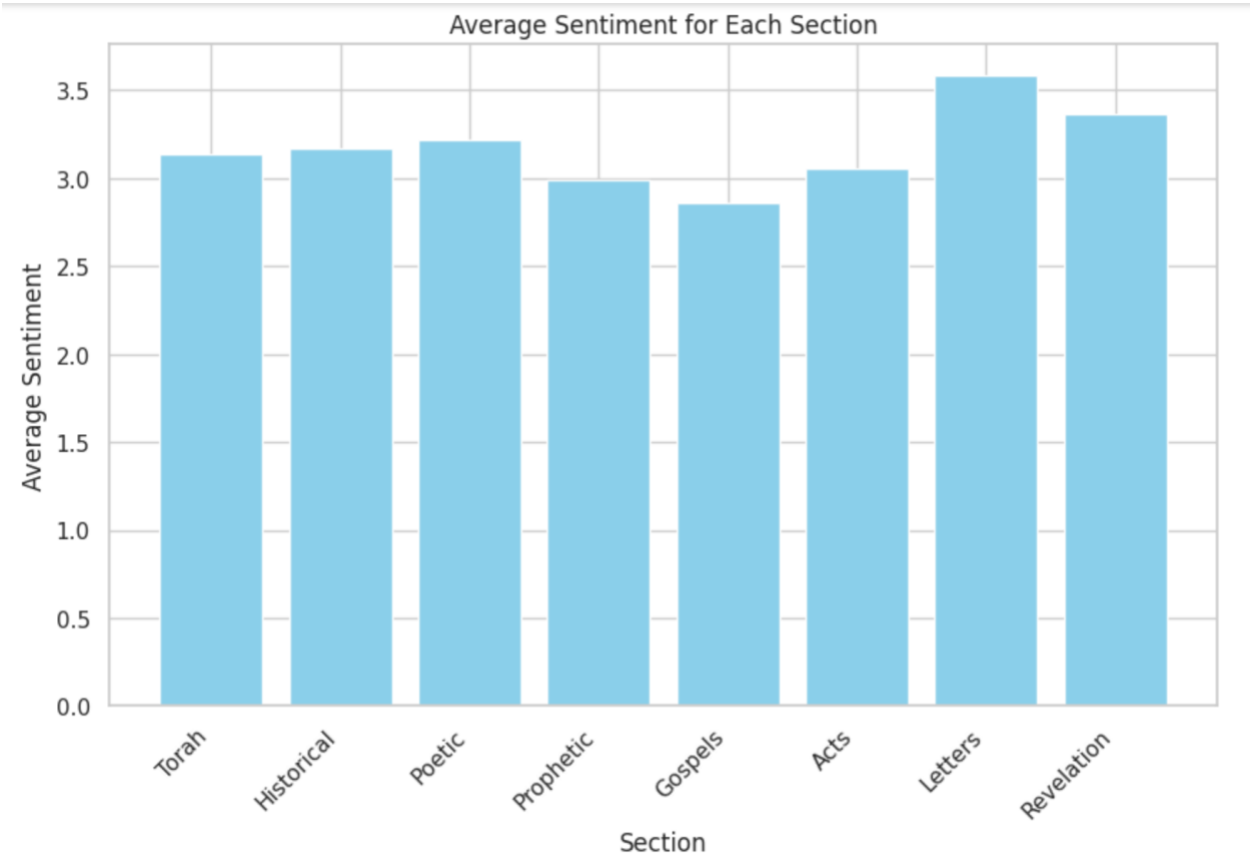
Arabic	2.689529
Dutch	3.397589
English	3.348703
French	3.725497
German	3.762873
Indonesian	1.395948
Italian	3.358191
Japanese	3.561578
Persian	3.409005
Spanish	3.166721
Turkish	3.013640

Since fine-tuning was not a possibility, and given the number of languages, length of text, and my knowledge of each language, it was not feasible to analyze all of the models more deeply (i.e. to do a kind of temperature reading). However we can look at the English results to get a sense at how the multilingual model performed. Some results seem to be fairly appropriate and accurate, while others are more random. For example, this verse had a sentiment of 5 (“Howbeit Jesus suffered him not, but saith unto him, ‘Go home to thy friends, and tell them how great things the Lord hath done for thee, and hath had compassion on thee.’”), and this verse had a sentiment of 1 (“When I heard, my belly trembled; my lips quivered at the voice: rottenness entered into my bones, and I trembled in myself, that I might rest in the day of trouble: when he cometh up unto the people, he will invade them with his troops”). At the same time, though, many high and low sentiment verses look like this (“And thou shalt take this rod in thine hand,

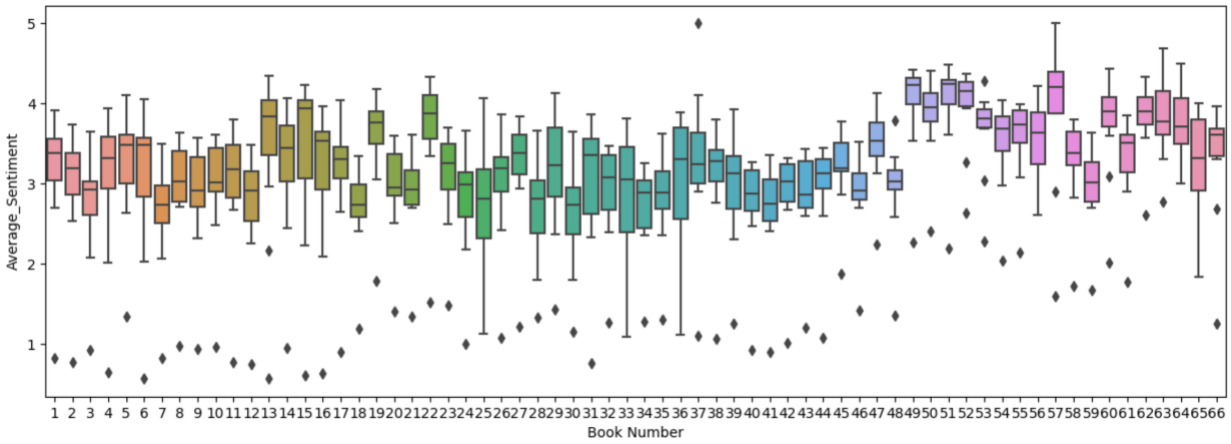
wherewith thou shalt do signs"; "And the twenty pillars thereof and their twenty sockets [shall be of] brass; the hooks of the pillars and their fillets [shall be of] silver.")

Disregarding the accuracy and still working with the results we got from the models, we can generate a few statistics to compare the sentiments across the languages. The following graph gives a picture of the average sentiment of each book and each broad section of the bible across all languages





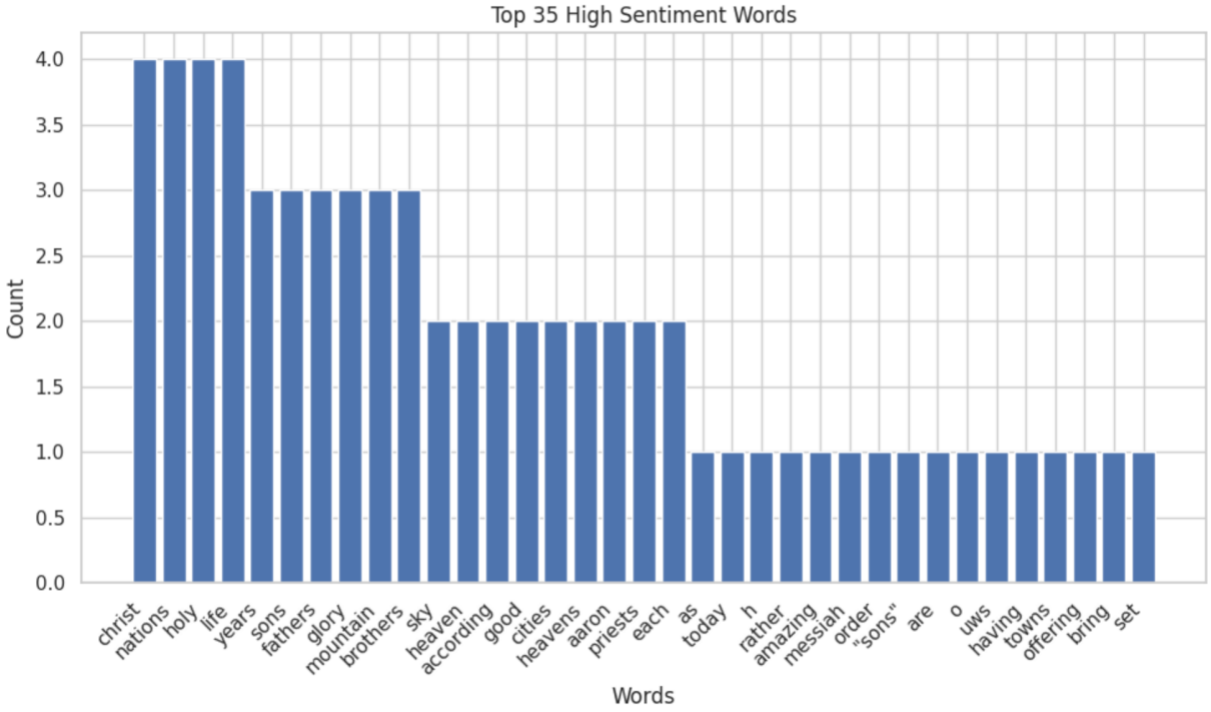
We can also utilize the standard deviation to see which books were scored more and less similarly across all languages.

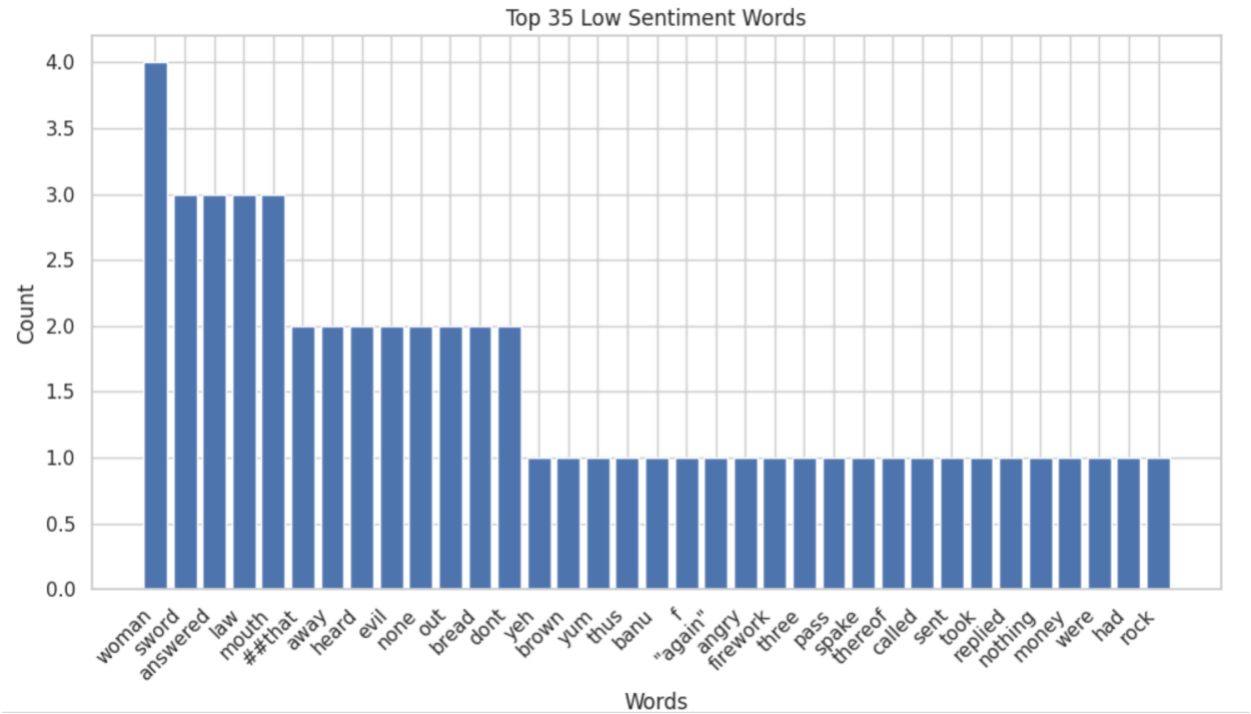


Across all texts, the book with the highest average sentiment was Colossians (with 4.1), and the lowest sentiment was Amos (with 2.6). The most consistently evaluated book was Acts of the Apostles, while the least consistent was Zephaniah, both with an average rating of 2.75.

Lastly, I wanted to see what words tended to appear in high and low sentiment verses. To compute this, I first tokenized the text of each translation, using a combination of different libraries (spacy, nltk, and huggingface models for languages not in either). Then I iterated through each token and language in order to get the counts of all words. After compiling the list, I translated each word via a Google Translate library (googletrans) and compared the results. The

results include the common words in (and exclusive to) high and low sentiment verses (across languages):





words within each language that are unique to the high and low sentiment verses:

Language	High sentiment words
Arabic	## h, as, but rather, for, for me, i, in order to, jesus, lord, the messiah, the sky, this is amazing, today, was, with it, you
Dutch	sons, are there, christ, father and, having, heaven, his, nations, o, saying, spirit, to make, towns, uws, years
English	according, bring, days, earth, fathers, good, great, heart, holy, jerusalem, judah, know, mine, offering, place, set, sons
French	aaron, altar, brothers, christ, cities, fathers, glory, heavens, jerusalem, life, moses, mountain, nations, servant, servants, spirit, time, years
German	brothers, give, hauses, heaven, life, nations, saint, servants, sons, soul, talked
Indonesian	big, good, heart, holy, is, jesus, life, love, place
Italian	aaronne, brothers, christ, fathers, given, glory, heart, here, infra, mountain, mr, nations, opera, priests, servants, sky, soul, spirit, taken, tribe
Japanese	a servant man, christ, david, descendants, give, given name, holy, mainly, many, miya, moses, please, road, see, ten, year
Persian	god, i will, the god, wi, ای, ن
Spanish	aaron, according, cities, construction site, darling, glory, heavens, holy, jacob, kingdom, kings, life, mercy, mountain, parents, priests, princes, siblings, soul, time, years
Turkish	david, day, each, later, like this, says, the one which, to you, together, until, verdi

Language	Low sentiment words
Arabic	## and that, ## f, ## yum, ##s, ##that, and he said, banu, brown, land, or, son, the people, then, thus, until, yeh
Dutch	again, angry, firework, made, no one, said, sword, thousand, three, to go, which, woman, you
English	answered, away, called, evil, heard, jesus, moses, neither, pass, say, sent, spake, therefore, thereof, took, two
French	flesh, had, law, money, mouth, no, none, nothing, path, replied, rock, said, say, sword, times, truth, were, wrong
German	answered, away, bread, face, fire, gold, hands, heard, jesus, out, person, prophet, sekel, spirit, sword, woman
Indonesian	## i, ## kah, ?, a, again, dead, dont, said, then, until, what
Italian	aveno, blood, came, conciossiachè, disso, even though, eyes, law, mouth, no, on the contrary, out, priest, saulle, spade, these, waterfall, why, woman
Japanese	anger, blood, body, chase, everything, evil, fire, if, juda, killing, take, thing, to tell, tsurugi, water, who, woman, yes
Persian	all, class, dont, israel, n, rejection, they
Spanish	altar, answered, blood, bread, cause, door, either, face, fire, hands, head, law, male, meat, mouth, priest, sin, stuff, water, waters, you have

Turkish	## yla, ##moment, ##that, jew, king of, more, no, opposite, place, sin, son, what

and words that are unique in each sentiment to the particular language:

Language	**Unique High Sentiment Words**
Arabic	‘##a', '##m', 'with it', 'he was', '## h', 'the messiah', 'for me', '## ha', '##na', 'on me', '##how much', '##dr', 'may be', 'machine', 'whose', '## wa', 'the earth', 'hey', '##r', '##t', '##k', 'because he', '## wen', 'the sky', 'today', 'the king', 'did not', 'as', 'this is amazing', 'for him', 'and all', '##b', 'in order to', '##h', '##they', '## j', 'but rather', '##yen'
Dutch	'country', 'saying', 'o', 'stuff', 'saying', 'uws', 'neither', 'to dawn', 'ulieden', 'done', 'father and', 'face', 'having', 'whole', 'leave', 'heeren', 'soil', 'sees', 'are there', 'middle', '"speaks"', 'to make', '"sons"', 'two', 'lord', 'vote', 'towns', 'uwer', 'come', 'said'
English	'behold', 'thy', 'us', 'put', 'upon', 'every', 'ye', 'hast', 'make', 'hath', 'shall', 'mine', 'saith', 'shalt', 'bring', 'take', 'forth', 'among', 'know', 'unto', 'thine', 'brought', 'offering', 'set', 'made', 'thee', 'thou', 'even', 'come'
French	'cause', 'dead', 'point', 'altar', 'fit', 'summer', 'fishing', 'eternal', 'have', 'blood', 'against', 'to do', 'medium'
German	'hauses', 'agree', 'a house', 'please refer', 'contrary', 'jehovas', 'did', 'languages', 'gods', 'happened', 'talked', 'water', 'speaks', 'volke', 'saint', 'should', 'language', 'cunt'
Indonesian	'has', '## an', 'which', 'there is', 'love', '## right', 'become', 'they', '##my', '##his', 'nation', '## lah', 'allah', 'so that', 'by', '##your'
Italian	'avea', 'therefore', 'sky', 'neither', 'tribe', 'eziandio', 'first name', 'boss', 'so since', 'infra', 'de', 'taken', 'im', 'down from', 'or', 'accui', 'opera', 'she said', 'gentleman', 'here', 'far', 'lor', 'aaronne', 'davide'
Japanese	'please', 'both', 'myself', 'many', 'descendants', 'sin', 'everyone', 'road', 'gata', 'moromoro', 'before', 'a servant man', 'eye', 'main', 'case', 'ten', 'while', 'aoi', 'bito', 'given name', 'miya', 'during', 'mainly', 'ground', 'signs', 'aho', 'year', 'alone', 'brother', 'look'

Language	**Unique Low Sentiment Words**
Arabic	'##a', '##m', 'he was', 'the people', 'brown', 'banu', '## ha', '##na', 'on me', '##how much', '## f', '##dr', 'may be', 'machine', 'whose', '## wa', 'the earth', 'hey', '##r', '##t', '##k', 'because he', '## wen', '## yum', 'the king', '## and that', 'did not', 'for him', 'and all', '##b', '##h', '##they', '## j', 'yeh', 'thus', 'and he said', '##yen'
Dutch	'country', 'firework', 'soul', 'saying', 'neither', 'see', 'to dawn', 'ulieden', 'done', 'whole', 'leave', 'said "', 'angry', 'heeren', 'sees', 'soil', 'no one', 'three', 'middle', 'speaks', '"lord"', 'again', 'vote', 'uwer', '"come"', 'to go', 'said'
English	'behold', 'thy', 'us', 'put', 'took', 'sent', 'upon', 'every', 'ye', 'spake', 'hast', 'make', 'pass', 'hath', 'shall', 'saith', 'shalt', 'forth', 'called', 'among', 'unto', 'thine', 'brought', 'thereof', 'give', 'thee', 'thou', 'even', 'come'

French	'were', 'money', 'fit', 'none', 'had', 'truth', 'point', 'replied', 'nothing', 'have', 'flesh', 'against', 'medium', 'times', 'big', 'to do', 'rock', 'wrong', 'summer', 'fishing', 'eternal', 'days'
German	'mr', 'happened', 'should', 'sekel', 'languages', 'gods', 'prophet', 'agree', 'a house', 'please refer', 'language', 'contrary', 'did', 'jehovas', 'kings', 'speaks', 'volke', 'given', 'cunt'
Indonesian	‘has', '## an', 'his', '## i', 'there is', '## right', 'become', 'again', '##my', '##his', 'nation', '## lah', 'allah', 'so that', 'by', '## kah', '##your'
Italian	'avea', 'saulle', 'aveno', 'these', 'eziandio', 'first name', 'boss', 'so since', 'why', 'de', 'im', 'down from', 'accui', 'conciossiachè', 'waterfall', 'disso', 'even though', 'she said', 'gentleman', 'the', 'far', 'lor', 'spade', 'davide', 'on the contrary'
Japanese	'anger', 'to tell', 'chase', 'myself', 'juda', 'everything', 'everyone', 'gata', 'moromoro', 'killing', 'before', 'eye', 'look', 'main', 'case', 'thing', 'tsurugi', 'while', 'aoi', 'bito', 'during', 'body', 'ground', 'who', 'signs', 'aho', 'yes', 'alone', 'brother', 'time', 'both', 'if'

Discussion

Looking at the results broadly, it is clear that my project only scratches the surface of this area of research. For one, the models are not adept enough to give highly accurate sentiment values of the biblical text. Many of the models tended to avoid more neutral sentiments (values of 2-4), which it arguably should not have done, given how many verses lack much of a strong sentiment (e.g. ‘The Lord said to Moses’). Being trained on modern use of language in non-literary contexts also makes it so that the models are only giving a *somewhat* accurate rating at best. It is a bit puzzling, then, why the Indonesian sentiment model gave such high accuracy scores, when its results were also the least consistent (at least in terms of overall average sentiment; the relative weights of sentiments across books are much more consistent). This lack of faith in the models makes it difficult to say whether these inconsistencies are statistically significant or not. Despite that, the results of my analysis still show promise for further research in this area.

Firstly, it is certain that fine-tuning the models would yield more robust results. Since the model could sometimes give fairly accurate sentiment values, it is likely that this could be extended to more of the text, if it is fine-tuned on a relevant dataset. It is also optimistic that much of the data was consistent. Across most of the translations/models, the Letters section had the highest average sentiment, while the preceding Gospels had the lowest average sentiment. Similarly, the Poetic section had the highest average sentiment for books in the Old Testament. These trends were even consistent in the Indonesian dataset (relatively). In general, I would expect that, at the book level, the relative average sentiment would carry over across languages. This is because there is no great difference in the content being described. Rather, I would expect variation to mostly occur in the kind of language/words used. I.e. how are particular words and ideas translated, and how do those changes alter the sentiment of the resulting text. Similarly, by analyzing areas of high or low sentiment, what are the linguistic features that carry that sentiment across?

Thus, I sought to get a measure of this by first tokenizing the datasets and analyzing the counts of words in high and low sentiment verses. Again, there are several flaws with this approach. For one, there are very few robust datasets for tokenizing the data (especially with the

non-European languages). Secondly, translating the words out of context (in order to compare cross-linguistically) can be wildly inaccurate. Words carry different meanings in different contexts, and sometimes sentiment can be conveyed through grammatical, not lexical, means (i.e. commands having a lower sentiment than polite requests). However, this approach was simpler, faster, and easier given my background, and it still gave some initial insights.

Across languages, it seems that 'God', 'king', 'Israel', 'children', 'people', 'Jesus', 'son', and 'people' often appear in high sentiment verses. However, these same words also often appear in low sentiment verses. Thus, it would seem helpful to filter out Bible-specific stop words. Factoring common words out leaves few high/low sentiment words that are in all translations, but some common ones include 'Christ', 'nations', 'holy', and 'life' for high sentiment verses, and 'woman', 'sword', 'answered', 'law', and 'mouth' for low sentiment verses. The two lists also contain words that are generally associated with positive or negative ideas, especially in a biblical and historical context. Otherwise the rest of the words are fairly neutral. Again, while the translation of the text only goes so far, it is clear that such an analysis can fruitful. I stopped myself from spending too much time analyzing these areas, since I would be basing my analysis on data that is not likely to be thorough and accurate. It was clear, given numerous errors in the resulting translations, that this method only has limited efficacy.

Digging a little bit deeper into the variance *across* languages is a bit more difficult. My initial approach was twofold. One, to look within each language for the words that only appear in high sentiment verses or low sentiment verses. Two, to look at which words appear exclusively in one language (both of which are displayed in the results section). Both of these would give me a sense of what words are unique in each language for positive and negative sentiment. It is difficult to understand the nature of the words without a full understanding of the given language and culture, but it is still clear that there *are* differences (i.e. whether 'evil' is the word most often associated with negative ideas in a given language, or instead if 'anger' is more typically associated with negative contexts in another language). So it is interesting for me to look through the results and notice the subtle differences within and across languages. Only with a better knowledge of each language could I make much more of the data.

In general it seems like my analysis is limited both by the resources in NLP and in how much computational linguistics can add to the broader field of linguistics. From a computational perspective, the models struggled to perform adequately on the given datasets. Furthermore, libraries and packages outside of English were heavily lacking, leading to questionable results. From a linguistic perspective, the current computational analysis lacks the language-specific contexts that would place the results into a broader narrative. Without additional knowledge on, say, Indonesian, I am not certain just how trustworthy the results of the model are. However, this project is a helpful start.

Taking this a bit further, my project ties into the conversation about multi-lingual NLP as well as translation studies. For one, much NLP research has centered around a handful of very similar languages, leading to a lack of models and resources in less-common and less-related languages. Secondly, while the field of translation continually makes immense strides, there are always additional dimensions, like sentiment, that translation models might not be attentive to. There are numerous ways to translate the same idea, and knowing which way to translate it depends on sentiment, among other features like formality, dialect, etc. This project has helped to highlight how investigating sentiment can be a viable way to contribute to both areas of research. That being said, recent research has suggested that more complex machine learning strategies could counteract several of these drawbacks.

Conclusion

This research project was as much an investigation into cross-linguistic NLP as it was a learning process in data science for me. Thus, much of the results is based solely on my initial dive into data science. Nevertheless, the goal of the project was to see what a cross-linguistic analysis of a singular text could reveal about the people and culture of the various languages. With the bible chosen as the main text, I looked into sentiment analysis in English, Spanish, German, Dutch, French, Italian, Arabic, Japanese, Turkish, Persian, and Indonesian. After running models to determine the sentiment of each verse in the respective translations, I found that the models, while largely inexact, gave a good idea as to the value in my approach. The models were generally consistent in labelling broad sections of the bible with a given sentiment, despite certain individual books having a wider range of consistency. Looking more closely into the language again revealed areas where the languages were both consistent and variable.

Given the areas where the research is lacking (tokenization libraries, accurate translation, sentiment analysis models) indicates areas where NLP is broadly lacking. Secondly, the value in cross-linguistic sentiment analysis factors into research and dialogues relating to machine translation and ways to most effectively translate content from one language to another. The ideas from this project can also be worked on for much longer in order to give a more thorough answer to my initial research question. A sentiment analysis model fine-tuned for historical/biblical text would be hugely beneficial in boosting the accuracy and implications of my research. Likewise, more thorough translations and language-specific knowledge would provide a better analysis of what the variances in my results say about each particular language and culture. Still, this exploration highlighted the importance of cross-linguistic NLP.

References:

Brooke, Julian, et al. "Cross-Linguistic Sentiment Analysis: From English to Spanish." *International Conference RANLP 2009*, pages 50–54. https://aclanthology.org/R09-1010.pdf.

Khalid, Amrita. "Spotify Is Going to Clone Podcasters' Voices -and Translate Them to Other Languages." The Verge, The Verge, 25 Sept. 2023, www.theverge.com/2023/9/25/23888009/spotify-podcast-translation-voice-replication-open-ai.

Matusov, Evgeny. "The Challenges of Using Neural Machine Translation for Literature." *Proceedings of the Qualities of Literary Machine Translation*, 2019, https://aclanthology.org/W19-7302.

Si, Chenglei, et al. "Sentiment Aware Neural Machine Translation." *Proceedings of the 6th Workshop on Asian Translation*, 2019, https://doi.org/10.18653/v1/d19-5227.

UNLV. "Chapter 15.2 Linguistic Reconstruction." *ALIC Analyzing Language in Context*, alic.sites.unlv.edu/chapter-15-2-linguistic-reconstruction/. Accessed 23 Dec. 2023.

Xu, Yuemei, et al. “A survey of cross-lingual sentiment analysis: Methodologies, models and evaluations.” *Data Science and Engineering*, vol. 7, no. 3, 2022, pp. 279–299, https://doi.org/10.1007/s41019-022-00187-3.