Analysis of Trends in Sermons from General Conference

Yvonne Andrewsen, David Morley, Adam Robertson

April 15, 2020

Abstract

Talks given by speakers in General Conferences of The Church of Jesus Christ of Latter-day Saints vary widely in subject and length. Therefore, they present interesting possibilities for bag-of-words analysis and for classification. In this paper, we use a variety of algorithms to answer the following questions: Can we classify talks from conference by any single feature (e.g. speaker, year, speaker's calling, etc.)? Can we measure any significant shift in the treatment of certain topics related to societal developments (including pornography and homosexuality) in talks over time? In our work, we show that classifying conference talks by calling is a fairly difficult problem, and none of our methods achieved an acceptable level of accuracy. Furthermore, we do not observe any significant shift in the use of keywords related to specific topics.

1 Introduction

1.1 Background

Members of the Church of Jesus Christ of Latter-day Saints gather every year in April and October in a worldwide Church conference, called General Conference. During these biannual conferences, senior Church leaders give counsel and instruction to members of the Church throughout the world. These messages are recorded and stored for later study on the Church's website General Conference. Speakers in Conference are generally selected from the leadership of the Church, which can be divided into the following categories: Prophet, First Presidency, Apostles, Relief Society General Presidency, Young Men General Presidency, Young Women General Presidency, Primary General Presidency, Sunday School General Presidency, and Seventy. Topics are not assigned; however, a speaker's message generally relates to their leadership assignment. In general, all members of the First Presidency and Quorum of the Twelve speak in every General Conference. Roughly seven or eight Seventies are also selected, along with one or two members from every other organization (Young Men, Young Women, etc.).

1.2 Research Questions

As the Church experiences growth, particularly international growth, the culture and needs of the Church change. These changes are reflected in General Conference. The themes discussed in conference vary over time. We are interested in analyzing these themes, and their changes over time, because this analysis may lead to insights into how the priorities of Church leaders change over time. This study considers the following questions:

1. What characteristics typically distinguish speakers from one another? (i.e. can we classify talks by speaker/calling)
2. How have themes from General Conference changed over time? In particular, how have themes related to societal developments (including homosexuality, pornography, and the role of women in the Church) changed over time?

1.3 Existing Research

Religious scholars have developed several tools that are useful for analyzing localized themes within conferences but are ill-suited for a longitudinal study. Most of the research in this field consists of microanalyses of themes within conferences. In 1986, Shepherd and Shepherd Gary Shepherd [1986] performed a systematic review of General Conference talks from 1830-1980. They randomly selected conference addresses to determine a list of themes, then charted the prevalence of these themes over time. Their study relied on human readers and they focused on the tone and audience for each theme. Their investigation was limited to a subset of conference talks. Our investigation is more comprehensive over the period from 1970 - 2019 because we analyze every talk. However, our method does not discriminate among the intended audiences of each talk nor the subtleties of each theme.

2 Data Collection

2.1 Web Scraper

We built a simple web scraper to grab each talk from every conference between April 1971 and October 2019 from the official website of the Church General Conference. Since we also wanted to include talks from 1970, we found and manually scraped text archives of the talks given in that year Conference Report. These scraping efforts resulted in raw text files, one file per talk, and an index table that contained the talk name, speaker, date, speaker's calling, and the corresponding text filename.

2.2 Data Cleaning

The different algorithms used required unique methods of cleaning ranging from just lowercasing the text to performing sophisticated algorithms before the data could be used. See individual subsections under Methods for specific cleaning methods.

2.3 Data Quality

Where possible, data was retrieved from official sources. When official sources did not exist, we checked references and otherwise did our best to determine the quality and authenticity of our sources. When scraping we checked the `robots.txt` file for each website and followed all prescribed guidelines. We declare that to the best of our ability, our data was ethically and honestly retrieved, processed, and stored.

2.4 Ethical Considerations

The primary ethical concern related to this project is the possibility that incorrect assumptions about the Church or its leaders could be drawn from our analysis. We emphatically state that our results and opinions are our own and do not represent, nor necessarily correspond with, the values and teachings of Church leaders. The conclusions we have made about the content of their messages should not be taken as a statement about any speaker's intent.

We also note that we are studying topics that have caused controversy for the Church. The results of our work may cause an appearance that the Church is shifting toward, or away from, an unpopular political opinion.

3 Methods

3.1 What are the best criteria to classify talks on?

Human readers familiar with the Church who are presented with the text of a recent or well-known talk are likely able to identify the speaker and the speaker's calling. We sought to discover the best criteria to use to train a computer to similarly classify talks. We performed the following process: encode the talks using a bag-of-words method, reduce the dimensions, and attempt a classification algorithm based on a talk's similarity to other speakers.

3.1.1 Bag-of-Words Method

The most straightforward way to encode a bag-of-words model is to use the counts of every word in a talk. After eliminating the common words with low semantic contribution (e.g. and, but, the, of, etc.), we encoded each talk as a row in a matrix where each column corresponded to the number of times a certain word was used. For example, the sentence "I love chocolate and I love math" would first become "love chocolate love math" and then be vectorized as:

love	chocolate	math
2	1	1

This method, however, was not satisfactory for our data set because it still gives a lot of weight to words common to many speakers. As a result, we further transformed our feature vector using a method called Term Frequency–Inverse Document Frequency (TF-IDF). This is a common method (see Volume 3, 10.2.3) that considers both the *term frequency* (the number of times a word is used in a talk (w_t) divided by the length of the talk (t)), and the *document frequency* (the number of times the word is used across all the talks in the corpus (w_c) divided by the total number of talks

(c)). Due to the size of the corpus, we took the log of the inverse document frequency to avoid overflow. Thus, for each word in each talk, we transformed the bag-of-words matrix by the rule $\text{tf-idf}_w = (w_t \cdot t) \log\left(\frac{c}{w_c+1}\right)$.

3.1.2 Using the Topical Guide

The first source of words for our bag-of-words method was the Church's Topical Guide -- a comprehensive list of topics of varying rarity, as well as their common synonyms. The entries in this guide were pared down so that every line was only one word long to serve as a suitable bag of words. A sparse weighted matrix was then created to represent the frequency of word occurrences for each individual talk.

With 3,855 talks and 3,950 words we were potentially checking for in each talk, as well as a great deal of correlation between words (for example, the words 'Jesus' and 'Christ'), we chose to reduce the matrix to only a few dimensions. For later analysis, we created 24 reduced representations of the talks, using the reduction techniques of PCA, tSNE, and UMAP, and reducing to 2-9 dimensions. From here, we had several different options to try and classify a speaker's calling given the words they used in their talk.

We first attempted a one vs. the rest classification approach. We selected one talk and did a binary classification on all other talks based on their similarity to the chosen talk. The binary classification was with respect to the chosen talk -- same calling or different calling as the one selected. After ordering all other talks based on their similarity to the chosen talk, we analyzed the utility of the classification by looking at the ROC and the area under that curve. The AUC for every talk using all 24 methods of dimension reduction was stored, and the spread was as follows:

mean	0.499444
std	0.018558
min	0.296138
25%	0.490356
50%	0.500157
75%	0.509688
max	0.660371

Table 1: Classification Results

Unfortunately, the mean and median areas were both less than 50%, indicating these to be poor classifiers. The best talk had an AUC of about 2/3 (['Nourished by the Good Word of God' by Daniel K Judd in the Sunday School Presidency), while the worst talk had an AUC of less than 1/3 ('Happiness Is Homemade' by LeGrand R. Curtis in the Young Men Presidency). Using a histogram to represent the spread, the curve appears extremely normal, and it would be reasonable to conclude that both our great classifiers and our poor classifiers perform differently due to chance, and not some underlying inherent differences in the speaker or their calling.

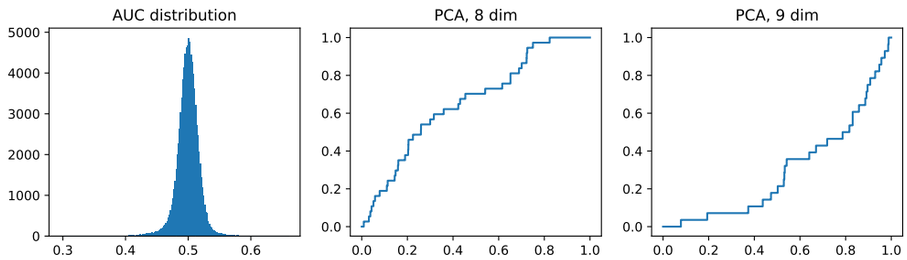


Figure 1: ROC Similarity comparisons

Next, we attempted to classify a talk based on the calling of the k Nearest Neighbors. Using a Pipeline to find the optimal way to measure difference and number of neighbors to consider, PCA regularly had a classification barely above .5 across all dimensions. UMAP and tSNE were both barely under .5. Unfortunately, neither of these are particularly impressive classifiers.

Rather than looking at kNN, we instead tried analyzing the clusters through KMeans. The downside to this is the clusters assigned are not immediately matched to an original calling they were intended to represent. Therefore, in order to provide some measure of analysis, for every cluster to which talks are assigned, we looked at the true callings of the talks. We assumed the most frequent true calling was the intended calling for the cluster. From there, we analyze what proportion of the cluster is from the mode category. The results here were erratic. Sometimes 100% of the talks assigned to a cluster were the same calling, other times the percent of speakers

belonging to the most frequent calling in the cluster was only 30%, with 20% being the absolute minimum possible value for 10 different callings.

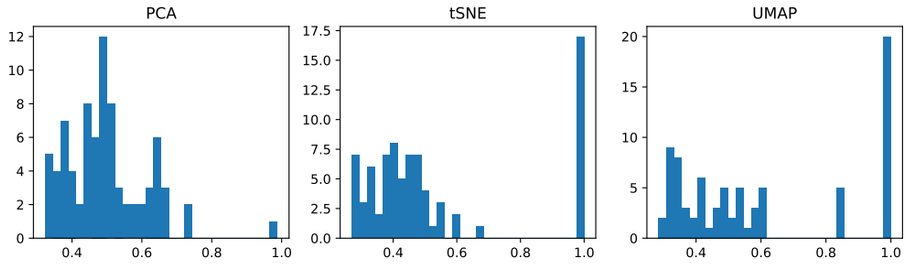


Figure 2: Proportion of majority calling within cluster

The results of these tests are all rather consistent -- the words a speaker uses cannot reliably help us to classify the calling of the speaker.

3.1.3 Using All the Words or Other Stylistic Features

To build upon results found using only words found in the Topical Guide, we also built a model using all information-contributing words. With the data formatted in this way, we were left with a feature matrix of 3,855 talks by 38,444 tf-idf transformed words. For comparison, we also generated a feature matrix to encode the stylistic choices of a certain speaker. This contained features like the length of the talk, the average number of syllables per word, and the reading difficulty.

First we tried to classify talks by calling as a preliminary step to identifying speakers. We implemented various clustering algorithms on both of our feature matrices.

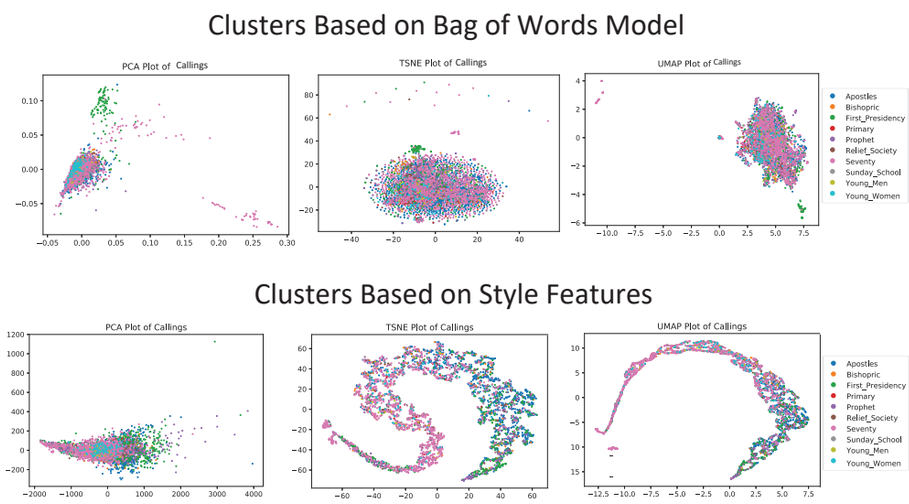


Figure 3: A comparison of different clustering algorithm and features.

As is apparent, this method was largely unsuccessful. Some small clusters appear, but none contain an entire class. Further work could show these clusters as belonging to individual speakers. We also attempted to cluster the talks based on the year they were given with similarly poor results. It is clear from this attempt that even with transforming the data to emphasize distinguishing words, there is no distinct predictive pattern to general conference speakers based on word choice and style. However, this finding does not hold when individual speakers are compared. For example, we were able to use the bag of words feature matrix to get up to 83% accuracy when classifying speakers. Here are the results of a comparison of two previous leaders of The Church of Jesus Christ of Latter-day Saints - President Gordon B. Hinckley and President Thomas S. Monson - with different algorithms.

3.2 Analyzing the Content of Talks Over Time

While the Church of Jesus Christ of Latter-day Saints is politically neutral, the Church is affected by political and societal developments. We analyze how Church leaders' usage of certain keywords related to developments that have occurred over the last fifty years has changed over time. In that time, the Church has faced pressure over its teachings regarding homosexuality, pornography, and the relationship between women and Priesthood authority in the Church. We use word embeddings to analyze the contexts in which keywords are used and then note how those contexts change over time.

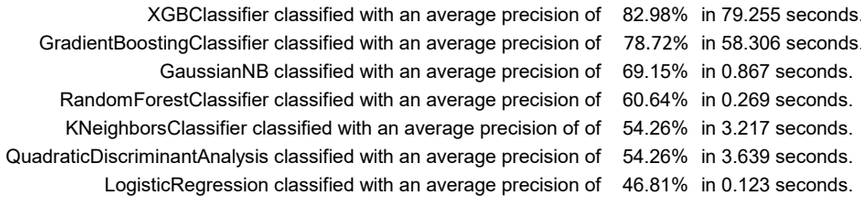


Figure 4: A comparison of different classifier algorithms with a bag-of-words model.

3.2.1 A Word on Word2Vec

The purpose of the Word2Vec algorithm is to create word embeddings. Given a corpus of text with minimal precleaning and a set of keyword, Word2Vec creates a many-dimensional space and assigns each keyword to a vector in that space. The distances between these word embeddings are crucial; when two words share a similar context, then their word embeddings are close in the vector space.

Word2Vec was originally developed and patented by a team of researchers at Google led by Tomas Mikolov Mikolov. There are two models that Word2Vec algorithms can implement to assign words to vectors: the continuous bag-of-words (CBOW) model or the continuous skip-gram model. Each model is a predictive one. CBOW predicts the current word given a set of context words, while skip-gram predicts the context of words given the current word. Both algorithms rely on a context window, which specifies the number of nearby words in each direction to consider relative to the current word (for example, if the context window has size 5, then the set of context words for the current word is the five words that precede the word in the corpus and the five words after). The authors of the algorithm recommend a context window size of around 5 for CBOW and 10 for skip-gram (https://code.google.com/archive/p/word2vec/). An advantage of the skip-gram model is that it takes into account the order of the surrounding words; context words that appear closer to the current word in the corpus are given more weight than distant context words. CBOW, on the other hand, considers the set of context words with the relative weights of each word being equal (so that the order in which they appear is not important). An advantage of the CBOW model is that it generally trains faster than skip-gram, especially when the corpus of text is very large.

3.2.2 A Word on WordNet

WordNet is a lexical database for the English language built and maintained at Princeton University. In addition to storing definitions and usage examples, WordNet contains synonyms for each word in the database. WordNet is accessible through an API in the `nltk.corpus` Python package.

3.2.3 Word2Vec on the Word of God

We ran Word2Vec on several different groups of talks, each of different sizes. We used the implementation of Word2Vec from the `gensim` package in Python, using both the CBOW and Skip-Gram models (each with a window of 5). When Word2Vec is run on the set of all talks from a single session of General Conference (October 2019), the corpus contains nearly 66,000 words. In this model, the words "Jesus" and "Christ" are given a score of 0.9995 by CBOW and 0.9992 by skip-gram. These scores make sense; "Jesus" and "Christ" are generally used interchangeably. Furthermore, the words "faith" and "repent", which we also expect to appear in generally similar contexts, are given a similarity score of 0.987 by CBOW and 0.999 by skip-gram. Table 2 contains additional cosine similarities among a sample of key words from the October 2019 conference.

Word 1	Word 2	CBOW Score	Skip-Gram Score
Baptism	Covenant	0.982	0.999
Repent	Repentance	0.981	0.999
Faith	Nelson	0.998	0.999
Gay	Sin	0.048	0.064
Gay	Repent	0.071	0.067
Pornography	Family	-0.108	-0.093
Pornography	Repent	-0.122	-0.096
Pornography	Sin	-0.131	-0.098

Table 2: Similarity Scores from the Oct. 2019 Conference

When Word2Vec models are trained on the entire corpus of talks from all of the sessions of Conference included in this study (1970-2019), which contains 8,266,933 total words, the CBOW model assigns the words ”faith” and ”repent” a cosine similarity score of 0.963, while the Skip-Gram model assigns them a score of 0.999. Both CBOW and Skip-Gram assign the words ”Jesus” and ”Christ” similarity scores above 0.999.

3.2.4 How Key Topics from Conference Change over Time

In order to analyze the similarities among a set of topics that were addressed in General Conference and are likely to have changed over time, we selected a set of keywords that represent these topics. We used these keywords, as well as a selection of their synonyms from WordNet, as the list of words that we ran through a Word2Vec model trained on the entire corpus of Conference talks. After reviewing the similarity scores among words belonging to the same theme, we removed several words from the dictionary to bring the average score for each theme above 0.5. After paring the set of synonyms down to an appropriate group, we created the following set of theme words (and their synonyms):

- Jesus (Christ, Savior, Redeemer)
- Family (Home)
- Marriage (Spouse)
- Homosexual (Gay)
- Priesthood
- Woman (Women)
- Pornography
- Sin (Wicked, Wickedness)
- Commandment (Teaching, Doctrine)
- Love

We measured the cosine similarity of each word against all other words in our list. We then took the average score of each word from one group against each word in another group to assign that pair of groups an average similarity score. The results are Figure 5 and Figure 6.

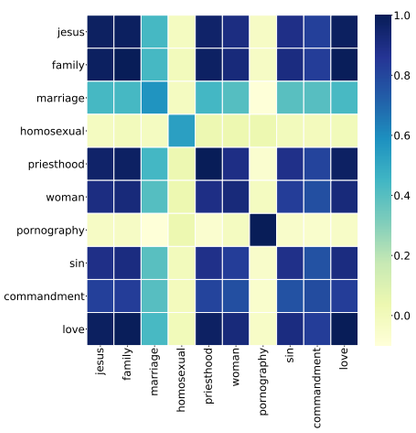


Figure 5: Average Scores from CBOW Model

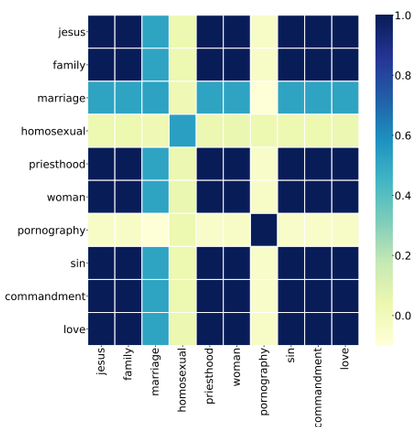


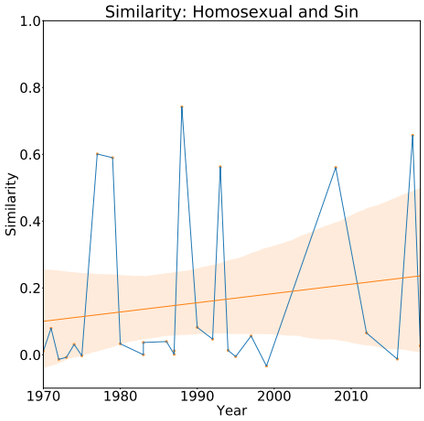
Figure 6: Average Scores from Skip-Gram Model

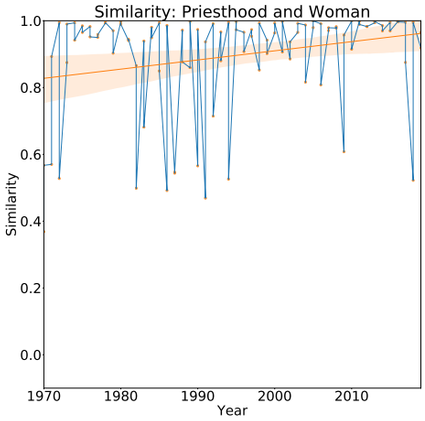
These models produced similar results. Some of the words selected for each theme had a much higher level of similarity in the model than others. For example, the average similarity among the words chosen for the theme ”Jesus” was 0.975 for CBOW and 0.999 for Skip-Gram. Meanwhile, the average similarity among the words chosen for ”Homosexual” was 0.535 for both CBOW and Skip-Gram. This disparity may be due to the fact that the word ”Jesus” is used much more frequently in Conference than the word ”Homosexual”, which only appears in 27% of Conference sessions.

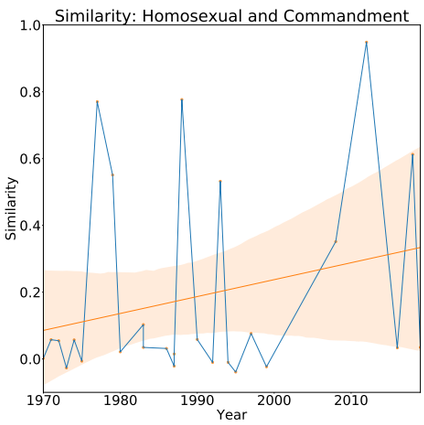
We also note that there seems to be a cluster of words that are all used in a similar context. The words ”Jesus”, ”Family”, ”Priesthood”, ”Woman”, ”Sin”, ”Commandment”, and ”Love” all score above 0.8 with each other in both models. This suggests that those words are all generally used together. The word ”Marriage” has a similarity score of roughly 0.5 with all of the themes listed above. Meanwhile, the words ”Homosexual” and ”Pornography” have very low similarity scores with all of the other words in the set.

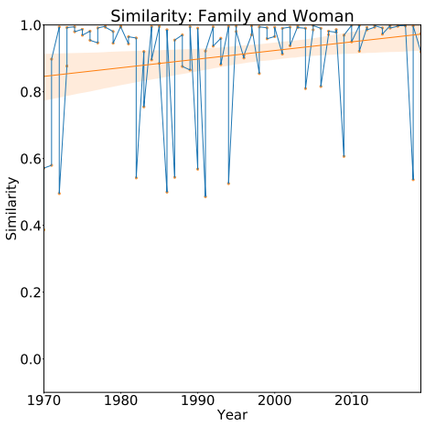
Finally, we performed Word2Vec analysis (using the CBOW model) on the text of talks from each conference individually, using the same theme words as before and taking their averages against all other words. We stored these similarities between each pair of theme words from each conference as a set of time-indexed data.

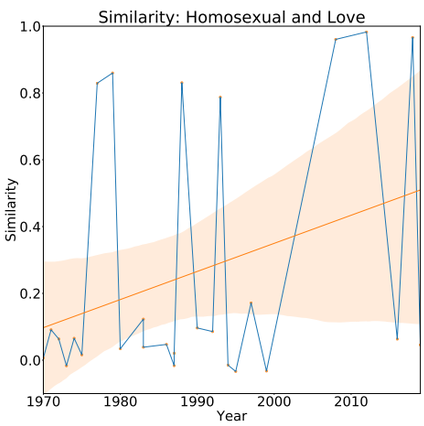
The following visualizations show how the similarities among various themes have changed over time. The orange line in each plot is a linear regression line, and the orange-shaded region is a 95% confidence interval.

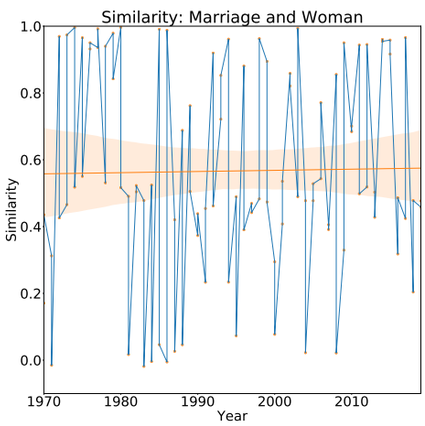












The data indicate that, over time, there does not seem to be very much correlation in a shift of context for most of our words. The plots above show that there is some correlation between the words ”Homosexual” and ”Love” over time. The data suggest that these two words may be used in similar contexts more frequently over time. A similar, though less significant correlation, is evident between the words ”Homosexual” and ”Commandment”, and even less of a correlation between ”Homosexual” and ”Sin”. There is some correlation between ”Woman” and ”Priesthood” and ”Family”, but no correlation between ”Woman” and ”Marriage”. Please see the Auxiliary File for the plots of the relationships among the other words. These data are not sufficient for us to conclude that there has been any significant shift in the usage of these words relative to one another over time.

3.3 Other Mentionable Work

In the process of analyzing General Conference talks, we used a significant number of algorithms and approaches, not all of which have been thoroughly discussed in this document. We chose to only include the most interesting and relevant methods in our discussion. Other methods were deemed unsuitable for various reasons. See Table 3 for specific examples.

Algorithm	Reason/Results
Topic Modeling with Latent Dirichlet Allocation (LDA)	Research Question not Pursued
Graph Kernel for Document Similarity Giannis Nikolentzos [2017]	Not Enough Time to Implement the Algorithm
Time Series	The data from the Word2Vec model are not correlated to their previous values. ARMA models produced from these data were not very accurate.
Random Forest	Tried to classify talks by the year when they were given. Accuracy was very low.
Gradient Boosted Classifier	Same as Random Forest.
XGBoost	Same as Random Forest.
CNN	Built a classifier that was only 45% accurate at classifying between talks by Russell M. Nelson and Dallin H. Oaks.

Table 3: Other algorithms used

4 Future Work

There is ample room for future work in classifying conference talks. We have attempted to classify by a speaker's calling. Additional work could classify talks by speaker using audio files of each talk. With this data, the anticipated accuracy of distinguishing speakers is very high. We hypothesize that our ability to predict the calling from the sound data would be about the same.

There is also room to expand on the results of the word embeddings discussed in this project. We evaluated the contexts of words from entire sessions of conference taken together. Future researchers may consider taking all of the talks from an individual speaker and analyzing them to see if the contexts in which that speaker uses certain words changes over time. Future work could also perform classification on the word embeddings. Word embeddings may be a more reliable classification feature for talks than the original text.

5 Conclusion

We conclude that the classification of talks from General Conference is a complex problem. Given that so many speakers overlap in word choice and in the content of their messages, it seems difficult for classifiers to accurately distinguish between different speakers or their respective callings. Furthermore, there is no data to indicate that Church leaders associate the word pairs that we studied any more or less over time. The set of Conference talks is an interesting data source, and additional research may draw value from them beyond what we have discovered.

References

Conference Report. Internet Archive, 2020. URL https://archive.org/details/conferencereport.

Gordon Shepherd Gary Shepherd. Modes of leader in the institutional development of mormonism. pages 125–126, 1986. URL https://doi.org/10.2307/3711457.

General Conference. The Church of Jesus Christ of Latter-day Saints, 2020. URL https://www.churchofjesuschrist.org/general-conference/?lang=eng.

François Rousseau Michalis Vazirgiannis Yannis Stavrakas Giannis Nikolentzos, Polykarpos Meladianos. Shortest-path graph kernels for document similarity. page 1890–1900, 2017. URL https://doi.org/10.18653/v1/D17-1202.

Tomas Mikolov. URL https://patents.google.com/patent/US9037464B1/en.

Princeton University. URL https://wordnet.princeton.edu/.