A Quantitative Analysis of Commencement Speeches

Abstract

Commencement speeches mark a critical transition point in the lives of graduates. The messages conveyed in these speeches offer advice for the future and calls to action, highlighting key values of society and the speakers at that moment. Unlike other types of speeches, commencement speeches have not been analyzed extensively as data using a quantitative approach. I study 825 commencement speeches given at 43 US colleges between 1890 and 2020 with natural language processing methods. The following analysis explores different measures of a speech, including its length, sentiment, and pronouns. My findings include that, over time, the speech length has decreased, and the general sentiment has fluctuated and shown some low points during challenging times in history, including the Second World War. Results related to the topics in the speeches include that mentions of civil rights increased over the decades and are related to whether the speaker uses primarily male or female pronouns.

1 Introduction

Commencement speeches give graduates some last pieces of advice for the next life phase. Commencement speakers are usually famous individuals who are considered to have had successful careers and lives. By nature, commencement speeches can convey important values in society at the moment when they were given. By analyzing commencement speeches, one can uncover social trends and explore what types of messages leaders choose to include (Rutherford 2004).

Unlike other types of speeches, such as presidential speeches, commencement speeches can be hard to find and therefore are less frequently analyzed. Although some colleges post past commencement speeches on their websites, most colleges make only a small subset of speeches publicly available on the web. There also exists a small online repository of commencement speeches given by female speakers at the Archives of Women's Political Communication. Most commencement speeches are stored in university archives or are not stored at all. Accessing archived speeches involves an often multi-week process of asking libraries to digitize and share the texts. Due to the limited availability of commencement speeches online, the number of analyses conducted on commencement speeches has been small. There are no articles in the prominent computer science or data science journals that have analyzed commencement speeches as data.[1] However, a few researchers have analyzed commencement speeches with a qualitative approach (i.e., Rutherford 2004, Bordelon 2010, Partch and Kinnier 2011, Bogdanowska-Jakubowska 2018), including sentiment, topics, and other measures. With the adoption of data science tools and natural language processing, however, a large number of speeches can be analyzed at once. This paper contributes to the existing literature by providing a novel quantitative approach to analyzing a large set of commencement speeches in a way that has not been possible with the solely qualitative analyses conducted previously.

The following research questions guided this work:

Q1: How does the length of a speech vary by year? Speech length is a common starting point for text analysis. For example, Statista provides an overview of inaugural speech lengths for all US presidents (*Length of inaugural addresses of U.S. Presidents 1789-2021* 2021).

Q2: How are pronouns (female versus male; 1st person, 2nd person and 3rd person pronouns) used in each speech? The topic of pronouns is especially interesting in the context of different college types. The data includes co-ed, religious, women's, historically Black, and other types of colleges. Pronouns may reflect important shifts over time as well as across different types of institutions.

Q3: What sentiment (positive, negative, neutral) is expressed in each speech? How does the sentiment change from the beginning to the end of a speech? The sentiment and emotions presented in a speech can be used to discover patterns across time and between types of institutions.

Q4: What are the most salient topics mentioned in commencement speeches? The collected commencement speeches span from 1890 to 2020. The past century has been marked by wars, conflicts, and moments of joy. The sentiment of a speech might reflect political issues that are especially relevant to college at a particular point in time.

Q5: How do the characteristics of a speaker, such as gender, relate to year, sentiment and pronoun use? Other studies have indicated differences in the pronouns a speaker used in their speech.

Q6: How do measures such as speech length, topics, and speaker characteristics, sentiment, and pronouns vary by type of college?

I will proceed in the following way: Section 2 provides a literature review on commencement speech analysis. Section 3 offers a brief overview of the technical details of text analysis, as well as natural language processing techniques used to analyze text data. Section 4 and 5 detail the data collection and processing steps. Section 6 explains the methods used to analyze the data in this paper. Section 7 details the overall findings for all colleges, and finally section 8 discusses the results, limitations, and future directions.

[1] A search for the terms "commencement speeches" and "commencement address" in the IEEE, ACM, and Springer online repositories did not yield any journal or conference papers on quantitative commencement speech analysis. The last search was conducted on May 8th, 2021.

2 Review of Literature on Speech Analysis

Speeches have been analyzed in great numbers, especially presidential speeches. For example, Alattar (2014) examines US presidential speeches to see how US presidents used linguistic acts and how that shaped what was being conveyed in each speech. The analysis used a linguistic approach to determine politicians' various methods to express their points. The authors found that the political landscape and events happening at the time of speech significantly shaped the tone and content of the speeches. The way that Alattar (2014) and others learn from analyzing speeches is the motivation for this paper.

In this section, I will first describe qualitative analyses of commencement speeches and then review quantitative analysis of speeches in general. Table 1 summarizes previous work on commencement speeches.

Author	**Number of speeches**	**Analysis Type**	**US Only**
Rutherford (2004)	171	qualitative	Yes
Bordelon (2010)	26	qualitative	Yes
Partch and Kinnier (2011)	90	qualitative	Yes
Bogdanowska-Jakubowska (2018)	100	qualitative	Yes
Zhu (2018)	60	qualitative	No
Bogdanowska-Jakubowska (2021)	87	qualitative	No
Huang (2021)	400	quantitative, blog post	Yes
This analysis	825	quantitative	Yes

Table 1: Overview of commencement speech literature

2.1 Qualitative analysis of US versus non-US commencement speeches

Zhu (2018) and Bogdanowska-Jakubowska (2021) compare commencement speeches across languages. Zhu (2018) compares English and Chinese commencement speeches from top universities in China and the United States. Their analysis focuses on how the speaker references the audience directly. They find that English speeches have more personal metadiscourse (a person discussing their own writing or thoughts) than Chinese speeches. Bogdanowska-Jakubowska (2021) compares commencement speeches given in the United States and Polish academic inauguration speeches, focusing on place, which includes situations, locations, and social aspects. There are common metaphors to describe the university and its role in a person's development in both the Polish inauguration speeches and the American commencement speeches. In general, the Polish speeches include fewer references to the speaker themselves or the audience. By contrast, the American speeches frequently mention the place the speech is being held and the speaker's background and story.

2.2 Qualitative analysis of US commencement speeches

Bogdanowska-Jakubowska (2018), Partch and Kinnier (2011), and Rutherford (2004) focus on commencement speeches given at US universities only. All three papers include speeches from a wide variety of US universities. Partch and Kinnier (2011) examined the content of 90 commencement speeches delivered between 1990 and 2000 at American universities. They found that the most important topics conveyed through the speeches included helping others, doing the right thing, expanding one's horizon, appreciating diversity, never giving up, cherishing others, and seeking balance. The topics of being true to oneself and cherishing others came up more often in speeches given at women's colleges and by female speakers. In addition, the authors found that women's colleges invited female speakers much more often than co-ed universities.

The speeches of particular individuals have also been the subject of study. Konfrst (2017) examined the commencement speeches given by US presidents. Konfrst showed that US presidents' commencement speech content depends on where they are in their presidency. If a speech was given within the first term of a presidency, the speech was largely used to promote their policy agendas. On the other hand, second-term commencement speeches focused more on building their legacy while pushing their policy agenda.

Similarly, Bogdanowska-Jakubowska (2018) uses Critical Discourse Analysis to examine the way speakers use their background as successful individuals and their fit as commencement speakers to focus on specific topics in their speeches. Their analysis focused on assessing how the topics of serving as a role model, past experiences, family, other influential individuals, achievements, dealing with failure, and mentioning the United States were present in each speech. They found that the beginning of speech is usually used for the speaker to address why they are qualified to be a commencement speaker. Most speeches also mention personal experiences of the speaker or belonging to a social group as framing for their speech. References to family come in different forms, including mentioning family members and their influence on the speaker. References to the United States also appear in speeches in various forms, including mentioning the Declaration of Independence, past presidents or essential individuals, discrimination and history, and American values.

By contrast, Rutherford (2004) focuses on the idea of moral choice and how the understanding of moral choice has shifted across the 20th century. Rutherford tested two hypotheses regarding how often the idea of choice was mentioned and how the idea of choice changed focus from collective to individual choice. According to Rutherford, the idea of choice appears more often and is interpreted in an individualistic manner over the 20th century. However, the idea of individualization is not always portrayed as solely positive. In general, Rutherford finds that speakers at women's colleges are no more likely to talk about career choices than at other colleges.

Bordelon (2010) took a closer look at commencement speeches at women's colleges and how they were part of identity formation for women during the early 20th century. Bordelon examines speeches by student speakers between 1910 and 1915 at Vassar College, which was then a women's college. The speeches given by graduating female students show how their education allowed this group of women to voice their opinions and shape the values of their community through the topics they addressed in their speeches. Often, the topics of choice included social justice and the advancement of women and were used as a platform to address further issues they deemed important.

2.3 Speeches Analyzed using Data Science Methods

The use of data science methods to analyze speeches is relatively new. No papers on the analysis of commencement speeches as data have been found in the top data science-related journals.[2] However, some papers have taken a data science approach to analyzing non-commencement speeches. Tucker, Capps, and Shamir (2020) use a dataset of approximately one million US congressional speeches and examine changes over time and differences between speeches given by Republicans versus Democrats. They find that women's identity is mentioned more often beginning in the 1980s. The readability index of the speeches changes over time, first increasing from a middle school to a high school level and then declining. The readability index also shows a partisan split, where Democrats have a higher readability index than Republicans. The vocabulary diversity also starts changing over time, and a partisan division is also found in this respect. The sentiment presented in the speeches also shows a growing positive sentiment present in speeches over time. However, when the opposite party had a president in the White House, the opposition party had more negative sentiment in their speeches.

Peignier and Zapata (2019) explore the discourse techniques used by Fidel Castro in his speeches. Their method involves looking at the word frequencies as well as topics that are conveyed in his speeches.

Although I have not found analyses of commencement speeches using a data science approach in peer-reviewed journals, some limited analyses have been conducted and presented as blog posts and GitHub repositories (Huang 2021, Abbaszadegan 2021, Rayapati 2021). Huang (2021) collected commencement speeches from the website Graduation Wisdom and FloyHub and analyzed both sentiment and topics of the speeches.

2.4 Methods Motivated by Literature Review

Like the qualitative work reviewed, I will examine sentiment by gender and look at topics mentioned in each speech. I will analyze sentiment and how the mention of specific topics in speeches has changed over time, similar to the approach taken by Alattar (2014). Many of the previous qualitative analyses focused on a specific group of topics. In this paper, the path to identifying topics will involve measuring bigram

[2] See Footnote 1

frequency, similar to the approach taken by Peignier and Zapata (2019). However, aspects such as readability and sentence length, as seen in Tucker, Capps, and Shamir (2020) will not be explored in this paper.

3 Review of Text Analysis Techniques

Multiple steps go into producing a text that a computer can understand. Not only does a text have to be in a format such as a text file or Word Document, but the text also needs to be transformed into features that can be processed and used to analyze sentiment or the topics in a text. This section offers a review of the steps that go into processing speeches and analyzing them.

3.1 Text Transcription

Historical texts or speeches are often not available in a digital format. Therefore, technologies such as Optical Character Recognition (OCR) can be used to digitize texts. OCR solutions are offered by Google, Amazon, and Microsoft. However, the accuracy of these technologies varies. Ughetta and Kernighan (2020) compared the performance of AWS Textract Google Cloud Platform's Vision and Microsoft Azure's Cognitive Services to that of human transcription. Their study used the Old Bailey corpus, containing over 180,000 pages of images of court records from 1674 to 1913, including text with demanding reading quality due to its age. The wide array of documents within the Old Bailey corpus makes it a useful dataset to benchmark the performance of various OCR technologies. Ughetta and Kernighan find that AWS had the lowest median error rate, while Azure had a faster transcription time.

Aside from written text, speeches can also come in spoken format, and studies such Papadopoulou, Zaretskaya, and Mitkov (2021) have looked at various speech-to-text transcription tools. For example, these authors compared various automatic speech recognition (ASR) systems, including that of Amazon, Microsoft, Trint, and Otter. They used videos from a lecture series given via Zoom and included videos from both native English speakers and non-native speakers. Each of the transcription results produced by the systems was processed with the help of the NLTK Python library. They looked at the post-editing effort, including how many words were changed, deleted, or inserted. Their study showed that AWS Transcribe performed poorly compared to other systems, including that of Microsoft and Trint.

In general, the performance of such systems may also be impacted by factors such as background noise or a speaker's intonation. Munot and Nenkova (2019) explored how emotions affect the performance of speech recognition systems. They used three English datasets of recordings of actors displaying different emotions. Munot and Nenkova found that emotions played a significant factor in recognition accuracy and that in comparing both improvised and non-improvised speech, the error rate was higher for improvised speech.

3.2 Analysis of Pronoun Usage in Speeches

Studies such as Lenard (2016), Newman et al. (2008) and Sendén, Sikström, and Lindholm (2015) have explored what pronoun patterns indicate about a speaker or writer of a text. The pronouns analysis in this paper was largely inspired by other studies examining pronoun usage (Lenard 2016, Newman et al. 2008). For example, Lenard (2016) examined congressional speeches and the different pronouns that are used in congressional speeches used by female and male speakers. Lenard showed that gender differences in pronoun usage were not statistically significant, by examining the number of times female and male speakers used certain pronouns. However, certain personal pronouns like "you" were used more by male politicians. In addition, female speakers focused more on their job rather than their personal life when using pronouns. A more general text and pronoun analysis are offered by Newman et al. (2008). In their research, Newman et al. (2008) compared and replicated findings from others, which indicate that female speakers use more pronouns than male speakers and longer sentences than male speakers. Sendén, Sikström, and Lindholm (2015) studied the use of the pronouns "she" and "he" in news articles written in 1996-1997 for Reuters. Using Latent Semantic Analysis (LSA), they found that pronoun "he" was used nine times more frequently than the pronoun "she". In addition, the usage of the pronoun "he" happened in a more favorable setting than the pronoun "she".

3.3 Sentiment Analysis

Sentiment analysis gauges whether a text has a negative, positive or neutral tone. Some sentiment analyses may also measure emotions such as anger, fear, and happiness. There are two general ways of approaching sentiment analysis. One is a machine learning approach (Alattar 2014) and another is a lexicon-based approach (Medhat, Hassan, and Korashy 2014). The machine learning approach focuses on a supervised learning and takes in labeled data to classify a text as positive or negative. Some of these methods include probabilistic models such as Naïve Bayes Classifier, Support Vector Machine (SVM), Decision Tree Classifiers, and Neural Networks (Medhat, Hassan, and Korashy 2014, Bonta and Janardhan 2019).

In contrast to the machine learning approach, a lexicon-based system with a pre-existing classification of words can also be used. Within the lexicon-based approach, there are dictionary-based approaches that use existing corpora, and the corpus-based approach that looks at the syntax and opinion words (Medhat, Hassan, and Korashy 2014). Some well-known lexicons include the LIWC (Linguistic Inquiry and Word Count) and GI (General Inquirer), which consist of a collection of words that are categorized as positive or negative. This binary classification approach does not account for the intensity of an expression and does not distinguish between something being good or great. However, valence-based lexicons do account for the intensity of sentiment. They take into account the emotions one associates with a particular word, such as the word “stress” being associated with a rather negative emotion. Valence scores for sentiment analysis are used in lexicons such as ANEW (Affective Norms for English Words), SentiWordNet, SenticNet, and VADER (for Valence Aware Dictionary for sEntiment Reasoning) (Bonta and Janardhan (2019), Ribeiro et al. (2016), Hutto and Gilbert (2014), Liu (2012)).

The ANEW lexicon includes ratings for 1,034 English words and scores the intensity of the sentiment from 1-9, where a valence score above 5 indicates a positive sentiment of different intensities. SentiWordNet includes 147,306 synsets (synonym sets) with three scores for neutral, positive, and negative words. Each score falls between the values 0 and 1. The SentiWordNet lexicon has the disadvantage that many synsets are mostly neutral and thus do not provide much information for sentiment intensity. SenticNet’s open-access lexicon includes 14,244 concepts such as adoration or wrath. For each concept there is a score value between -1 and 1 (Gonçalves et al. 2013, Ribeiro et al. 2016).

VADER uses a lexicon constructed via basic lexicons, including LIWC, ANEW, and GI, and includes over 9,000 lexical features and emoticons. VADER was constructed with the help of Amazon Turk (MTurk), where ten independent human raters were asked to rate a word on a scale from extremely negative to extremely positive (from -4 to 4). In total, the creators of VADER were able to collect a list of 7,500 lexical features with sentiment intensity scores, as well as whether each word is positive or negative (Hutto and Gilbert 2014). In addition, they also incorporated punctuation (such as exclamation marks), capitalization, negation, and degree modifiers when computing intensity. VADER has performed particularly well in comparison to other methods, as Hutto and Gilbert (2014) showed in their benchmark study, comparing the performance of VADER to other lexicons and machine learning approaches on social media posts, newspaper articles, and movie and product reviews.

3.4 Topic Modeling

The process of extracting topics from a text is called topic modeling. There are various methods of topic modeling, including Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Analysis (PLSA), and Latent Dirichlet Allocation (LDA) (Barde and Bainwad 2017). LSA (Latent Semantic Analysis) is a form of factor analysis and uses Singular Value Decomposition (SVD) to find topics. It does so by utilizing SVD to re-express highly correlated feature variables by new unrelated components. However, the potential downsides of these methods include that the factors do not have an intuitive explanation, and the direct relationship between individual factors and topics is difficult to inspect.

Another method is non-negative matrix factorization (NMF). This method relies on the fact that topics contain inherent clustering. It is set up as an optimization problem and factors original matrix to minimize loss and uses non-negativity constraints on mapping from topics to words. Both NMF and LSA use linear algebra to build low-rank approximations of the document-word matrix to find optimal representation.

In contrast to these matrix methods, Latent Dirichlet Allocation (LDA) tries to find out what topics a given document most likely belongs to and how documents might have been generated given a set of topics.

LDA is a probabilistic and generative method, and there also exist various variations of LDA, which is a special variation of probabilistic LSA (PLSA).

Another method is Latent Semantic Indexing (LSI) which uses indexing and information retrieval methods. By using SVD, LSI can identify the relationship between terms and concepts in a collection of documents and determine which terms appear in the same context.

Another approach is to use n-grams (contiguous sequences of n words) to identify topics within a text. Nokel and Loukachevitch (2015) identifies bigrams and use the bigrams with the highest frequency to create topics. By incorporating bigrams in PLSA, the researchers were able to improve the quality of the topics created. Similarly, Velcin, Roche, and Poncelet 2016 also use bigrams when creating topics for two specific cases, a set of abstracts from scientific papers and a set of news articles. For both datasets, extracting bigrams allowed for the fairly accurate creation of topics.

More advanced methods, including using BERT, a pre-trained bidirectional language model, have worked well with various natural language processing tasks. For example, Peinelt, Nguyen, and Liakata 2020 present a method of using BERT to predict semantic similarity and identification of certain topics.

3.5 Methods Used in this Study

I used AWS Textract for text extraction due to the average high transcription accuracy rate. The choice of AWS Textract was made based on the low costs of using AWS and the ease of creating data pipelines. Although studies such as Papadopoulou, Zaretskaya, and Mitkov (2021) recommended other systems, including Google and Trint's speech-to-text implementations, AWS Transcribe was used as the choice for transcribing spoken speeches. The choice of using AWS Transcribe was primarily driven by the fact that the speeches were stored in S3 buckets on the AWS platform. The AWS platform offered an easy way to combine the transcription of pdfs and mp4 files that included spoken speech. The advantage of AWS Transcribe is that it also returns the confidence for each transcribed word. Because of the ease of its usage and the relatively high confidence in most of the spoken speeches, it was used on, AWS Transcribe was selected.

The built-in NLTK stopword lexicon as well as the Porter stemmer were used as part of the text preprocessing step. Because of the simplicity of Norvig's spellchecker (Norvig 2021) implementation, the TextBlob implementation of Norvig's spellchecker (Loria 2018) was used to process the commencement speeches before analyzing them. VADER was chosen as the method for sentiment classification in this paper because of its simplicity and fast computation time. Even though many forms of extracting topics exist, I focused on a bigram-based approach due to its simplicity and straightforward comparisons with other variables in my collected dataset.

4 Data Collection

4.1 Data Collection Approach

Commencement speeches are frequently stored in university archives as written documents, video, or audio. To gain access to a wide array of speeches, I reached out to over 120 university and college archives. The initial list of colleges was intended to reflect the broad landscape of different types of institutions, to allow for comparisons between school types. The colleges on the initial list fall into several categories, including single-sex institutions, historically Black colleges and universities (HBCUs), religious institutions, liberal arts colleges, and large universities.

In the 2017-2018 academic year, the latest year for which information was made available, there were a total of 6,502 postgraduate institutions in the United States, of which 2,828 were 4-year colleges (*National Center for Education Statistics* 2021).

In 2018, there were 101 HBCUs in the United States (*National Center for Education Statistics* 2021). Making up roughly 3.6 % of all 4-year colleges in the United States. Similarly, there are 37 women's colleges in the United States as of 2021 (*Find a College Search — Women's College Coalition* 2021), making up around 1.3 % of all 4-year colleges. There were 772 public and 1,907 private 4-year colleges in the United States in the academic year 2019-2020 (*Digest of Education Statistics, 2016* n.d.). To learn about each type of institution, I wanted to make sure that the dataset included each college type.

The US News College List (*The Best National Universities in America* 2021) was used to compile a list of colleges to contact. The focus was on selecting colleges that were considered to be in the top 100 institutions in their respective categories. This decision was made with the idea that schools appearing higher in the rankings usually have more funding and thus would likely have archives of their commencement speeches. In addition, the list of colleges used by Partch and Kinnier (2011) was used to broaden the original list of colleges that were reached out to, since I knew that those schools had been able to provide speeches to those authors.

4.2 Overview of Collected Data

Overall, 932 speeches from 44 colleges across the United States were collected. However, some of these were only summaries of speeches or excerpts and thus were not included in the final dataset that was analyzed. Speeches from Dartmouth College were removed due to issues digitizing the speeches. The final dataset that was analyzed included 825 speeches from 43 colleges. The 43 colleges represent 18 states in the United States, 32 co-educational institutions (including Vassar College, which used to be a women's college), ten women's colleges, and one men's college.[3] There is one college that is an HBCU, all of the seven sisters colleges, 2 Ivy League colleges, 12 liberal arts colleges (not including seven sisters), and three conservative institutions. The conservative classification was determined based on rankings for the most conservative colleges in the United States (Berkman 2021), as well as the self-description of these institutions on their web pages. A more detailed overview of the data collection process can be found in the Appendix (Figure 18 shows the data collection process and how the 825 speeches that are part of this analysis were collected).

There were a total of 10 public universities and 33 private universities included in the dataset. The ratio of private to public universities roughly mirrors the distribution currently present in the United States, as seen in the statistics presented by the National Center for Education Statistics (*Digest of Education Statistics, 2016* n.d.). In addition, five of the colleges currently have a religious affiliation. Many of the other colleges were once affiliated with a particular religion but became secular.

The colleges in the final dataset are a subset of the colleges on the initial list. Due to the COVID pandemic, many college archives were closed during the data collection phase and were unable to share their speeches. This issue and the fact that not all colleges archive commencement speeches limited the number of speeches from certain types of colleges. There are currently only four all-male colleges in the United States, and I received speeches from one. Similarly, I received speeches from one HBCU.

Figure 1 shows the college and years of speeches that are included in the dataset. The colleges are ordered by type, with the women's colleges at the very top, followed by the men's college institutions and, finally, co-ed schools. St. John's University, the only men's college, is a small liberal arts institution with both female and male speakers. In the analysis, each college was only assigned to one college type category, even if they might have fallen into multiple categories. For example, Smith College was labeled a the seven sisters colleges rather than a liberal arts college. Table 2 shows which category each college falls under.

The gender of the speaker was coded as either male or female. In many cases, the speaker's gender was inferred by their name or via an Internet search. There may be speakers that fall outside the gender binary, but the gender variable was treated as binary in this paper. In addition to the text speech and name of the speaker, I collected the year of the speech, the college name, the occupation of the speaker, and the speech title. The college's location, size, type, public/private status, and any (religious) affiliation were determined by online searches.

Colleges with fewer than 3,000 students were determined to be small institutions, while a middle size institution was defined as a college with anywhere between 3,000 and 15,000 students, and a large institution was one with a student population of 15,000 or above.

5 Data Processing

Commencement speeches came in various formats, including PDFs, Word documents, video, and audio. All have to be converted to text that could be easily processed.

[3] The type of college is determined by the current student population a college serves. Therefore, Wells College, which turned co-ed in 2004, is part of the liberal arts group instead of the group of women's colleges.

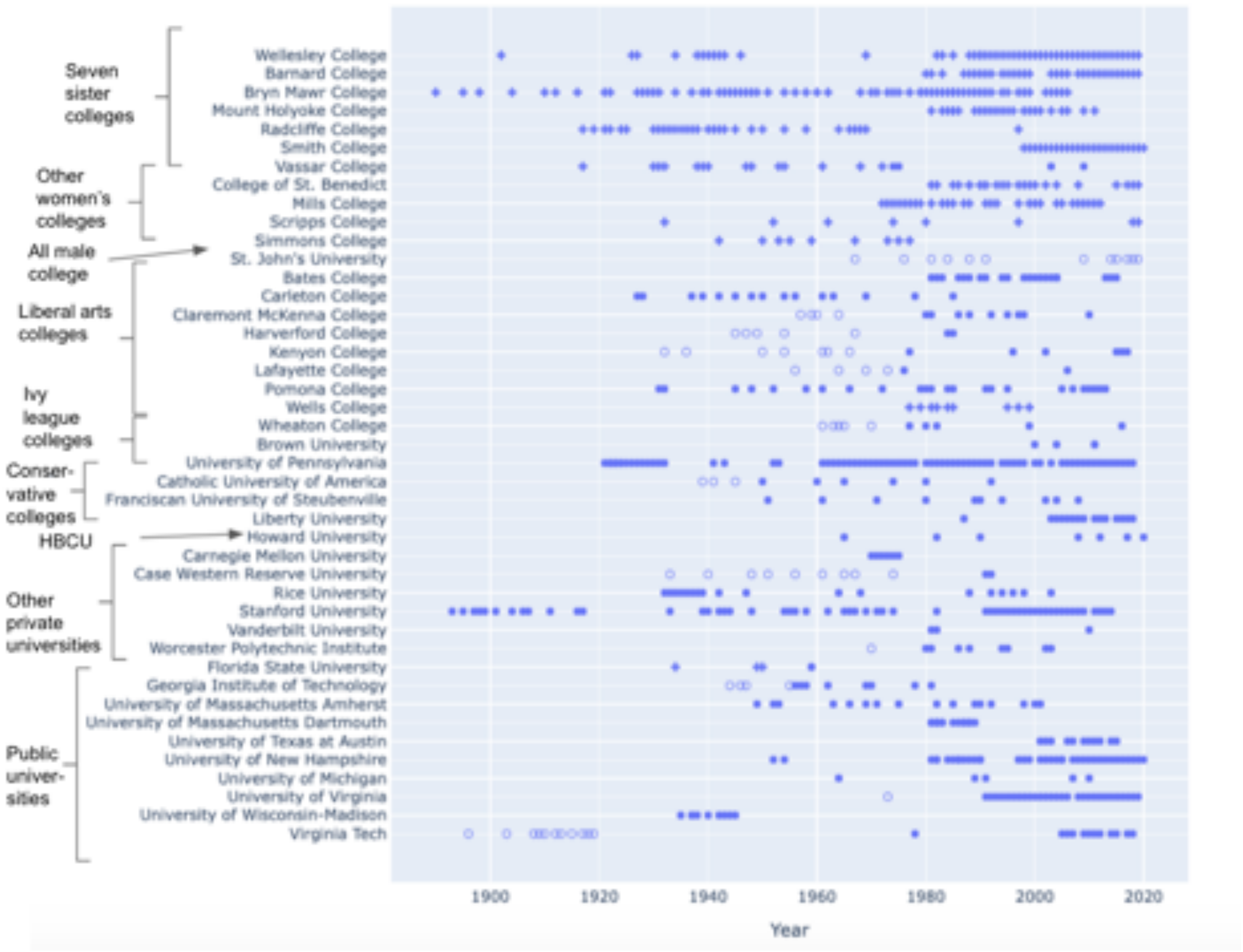


Figure 1: Speeches collected for different colleges with the symbol indicating if the college is a single-sex or co-ed institution when a particular speech was given. The plus signs indicate an all-female institution, the filled-in circles indicate co-educational institutions, and the empty circles indicate all-male institutions.

College Affiliation	**Colleges**
Conservative	Catholic University of America, Franciscan University of Steubenville, Liberty University
HBCU	Howard University
Ivy League	Brown University, University of Pennsylvania
Liberal Arts	Bates College, Carleton College, Claremont McKenna College, College of St. Benedict, Haverford College, Kenyon College, Lafayette College, Mills College, Pomona College, Scripps College, St. John's University, Wells College, Wheaton College
Seven Sisters	Barnard College, Bryn Mawr College, Mount Holyoke College, Radcliffe College, Smith College, Wellesley College, Vassar College
Other	Carnegie Mellon University, Case Western University, Florida State University, Georgia Institute of Technology, Rice University, Stanford University, Simmons College, University of Massachusetts Amherst, University of Massachusetts Dartmouth, University of Texas at Austin, University of New Hampshire, University of Michigan, University of Virginia, University of Wisconsin-Madison, Vanderbilt University, Worcester Polytechnic University, Virginia Tech

Table 2: Categorization of each college.

Table 3 shows the breakdown of different formats that speeches came in before being processed.

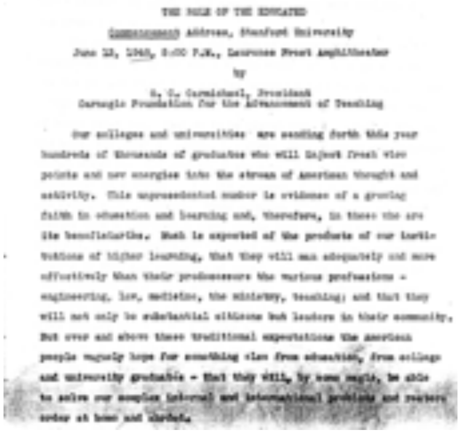
Format type	**Number of speeches**
PDF	637
Video	15
Word Document	14
Scraped from the web	159
Total number of speeches	825

Table 3: Breakdown of original speech formats

5.1 Speech Transcription

The speeches need to be first transcribed into a format that can be used for further processing. The speeches that were directly scraped from the web proved to be the easiest to work with. For example, Wellesley College has a list of some of the past commencement speakers and their speeches on their website (*Commencement*

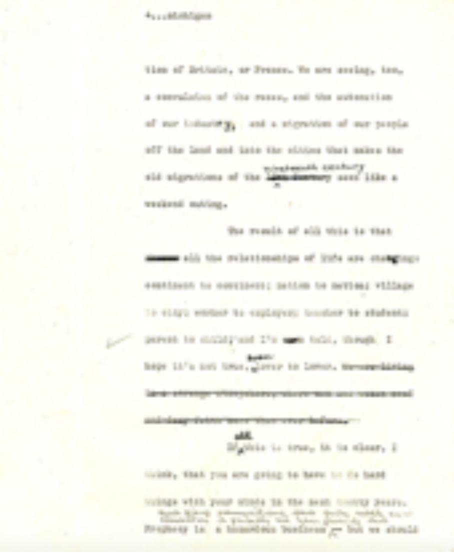
Archives — Wellesley College 2021). The available commencement speeches were scraped from this website using Selenium, a Python library. The speeches were saved in text files that could be easily read into a table for further processing. Speeches that were not in text format were transcribed with Amazon Web Services Textract (AWS Textract), and speeches in video format were transcribed using AWS Transcribe, as explained earlier.



THE ROLE OF THE EDUCATED
Commencement Address, Stanford University
June 13, 1948, 8:00 P.M., Laurence Frost Amphitheater
by
O. C. Carmichael, President
Carnegie Foundation for the Advancement of Teaching

Our colleges and universities are sending forth this year hundreds of thousands of graduates who will inject fresh view points and new energies into the stream of American thought and activity. This unprecedented number is evidence of a growing faith in education and learning and, therefore, in those who are its beneficiaries. Much is expected of the products of our institutions of higher learning, that they will man adequately and more effectively than their predecessors the various professions - engineering, law, medicine, the ministry, teaching; and that they will not only be substantial citizens but leaders in their community. But over and above these traditional expectations the American people vaguely hope for something else from education, from college and university graduates - that they will, by some magic, be able to solve our complex internal and international problems and restore order at home and abroad.

Figure 2: Example of a speech with smudges. First page of the speech by O. C. Carmicheal given at Stanford's 1948 commencement



4...Michigan

tion of Britain, or France. We are seeing, too, a revolution of the races, and the automation of our industry, and a migration of our people off the land and into the cities that makes the old migrations of the ~~[illegible]~~ nineteenth century seem like a weekend outing.

The result of all this is that [illegible] all the relationships of life are changing: continent to continent; nation to nation; village to city; worker to employer; teacher to students; parent to child; and I'm told, though I hope it's not true, [illegible] lover to lover. ~~[illegible]~~

If this is true, as is clear, I think, that you are going to have to do hard things with your minds in the next twenty years. [illegible] Prophecy is a hazardous business — but we should

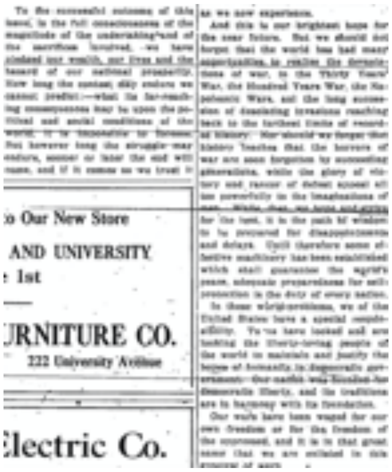
Figure 3: Example of a speech that was difficult to transcribe. One of the pages of the speech by James B. Reston given at the University of Michigan's 1965 commencement

However, Amazon Textract was not always able to successfully process PDF files that had slightly faded text (see Figure 2), thus producing erroneous text transcriptions. In addition, other speeches having crossed out text or unrelated texts as shown in Figure 3 complicated the transcription and cleaning process. Some texts also contained text in multiple columns or unrelated speeches or articles in one document. Figure 4 shows an example of a speech that not only includes two columns but also advertisements and streaks throughout the text. Therefore, after the speeches were collected, a manual review of all speeches was conducted to note which ones needed additional processing. Although AWS Textract has a way of extracting text from documents with multiple columns, because the text oftentimes was not cleanly formatted and contained multiple streaks as seen in Figure 4, a manual adjustment of the documents had to precede the text transcription. For all of the speeches that were hard to process solely via AWS Textract, I manually created a document with only one-column text, as seen in Figure 5.

Aside from AWS Textract, other services allow for text transcription, including the Google Vision API. To choose between the two different products, a handful of pdfs were selected to compare the transcription quality produced by the Google Vision API and AWS Textract. In both cases, using different spell checkers, the number of misspelled words was roughly the same. Therefore, AWS Textract was used to process the rest of the speeches as it offered an easy pipeline for transcription.

5.2 Spell Checking

Because the error rate of AWS Textract was relatively high for certain speeches, before further processing the speeches, a spell checker had to be applied. A comparative analysis of two spell checkers in Python was conducted. The performance of spellchecker module and the TextBlob implementation of a spell checker were evaluated by trying them on a set of five speeches. The choice of performance was made after reading through the spell-checked text from each spell checker and deciding on which version produced the more readable text. Figure 6 shows a snapshot of the spell checked version of one of the speeches. The small



To the successful outcome of this issue, in the full consciousness of the magnitude of the undertaking and of the sacrifices involved, we have pledged our wealth, our lives and the hazard of our national prosperity. How long the contest may endure we cannot predict,—what its far-reaching consequences may be upon the political and social conditions of the world, it is impossible to foresee. But however long the struggle may endure, sooner or later the end will come, and if it comes as we trust it

o Our New Store

AND UNIVERSITY

1st

URNITURE CO.

222 University Avenue

lectric Co.

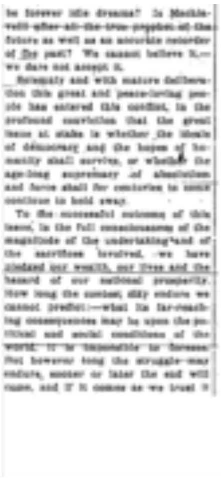
as we now experience.

And this is our brightest hope for the near future. But we should not forget that the world has had many opportunities to realize the devastations of war, in the Thirty Years' War, the Hundred Years War, the Napoleonic Wars, and the long succession of desolating invasions reaching back to the farthest limits of recorded history. Nor should we forget that history teaches that the horrors of war are soon forgotten by succeeding generations, while the glory of victory and rancor of defeat appeal all too powerfully to the imaginations of men. While, then, we hope and strive for the best, it is the path of wisdom to be prepared for disappointments and delays. Until therefore some effective machinery has been established which shall guarantee the world's peace, adequate preparedness for self-protection is the duty of every nation.

In these world-problems, we of the United States have a special responsibility. To us have looked and are looking the liberty-loving people of the world to maintain and justify the hopes of humanity in democratic government. Our nation was founded for democratic liberty, and its traditions are in harmony with its foundation.

Our wars have been waged for our own freedom or for the freedom of the oppressed, and it is in that great cause that we are enlisted in this struggle of ours.

Figure 4: A snapshot of John M. Stillman 1917 Stanford commencement speech that was difficult to transcribe due to the multiple columns and unrelated information on the page.



be forever idle dreams? Is Machiavelli after all the true prophet of the future as well as an accurate recorder of the past? We cannot believe it,—we dare not accept it.

Solemnly and with mature deliberation this great and peace-loving people has entered this conflict, in the profound conviction that the great issue at stake is whether the ideals of democracy and the hopes of humanity shall survive, or whether the age-long supremacy of absolutism and force shall for centuries to come continue to hold sway.

To the successful outcome of this issue, in the full consciousness of the magnitude of the undertaking and of the sacrifices involved, we have pledged our wealth, our lives and the hazard of our national prosperity. How long the contest may endure we cannot predict,—what its far-reaching consequences may be upon the political and social conditions of the world, it is impossible to foresee. But however long the struggle may endure, sooner or later the end will come, and if it comes as we trust it

Figure 5: Example of how the speech by John M. Stillman looked after the manual processing of the speech.

study indicated that the TextBlob implementation module did a better job correcting spelling mistakes and thus providing a more readable text. Therefore, the TextBlob implementation was used for processing the speech texts after they were transcribed via AWS Textract.

6 Methodological Choices for Data Analysis

Inspired by the research questions, this section will describe what choices were made to study pronouns, sentiment, and topics in the collected commencement speeches.

6.1 Evaluating Pronoun Usage

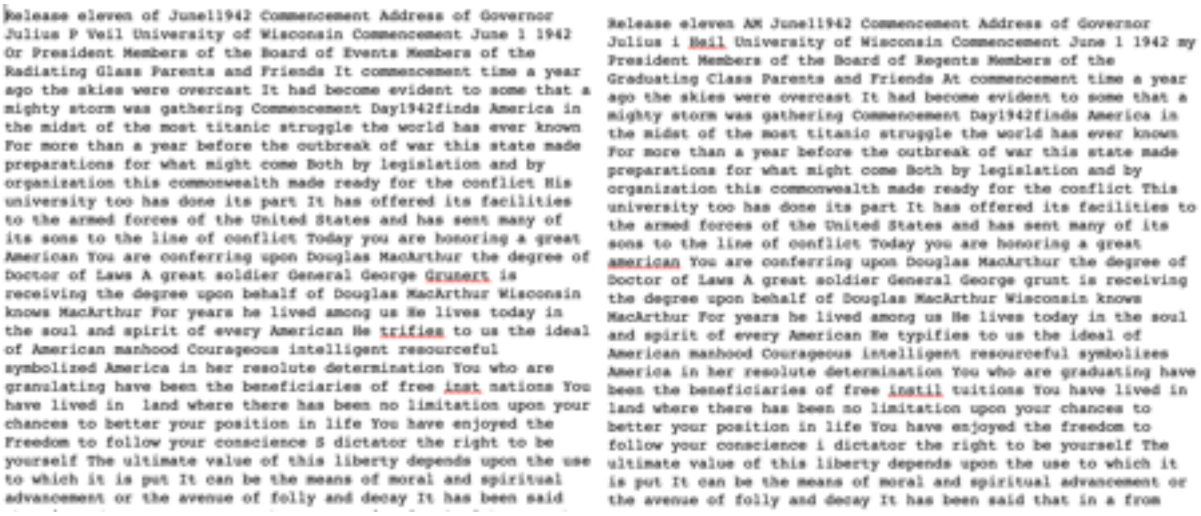
As discussed in Section 3, previous authors have analyzed the use of pronouns in various contexts to learn about a speaker's attitudes and perspectives (see Lenard 2016, Newman et al. 2008). The pronouns used in this paper were largely inspired by previous research done on the use of gendered pronouns in texts (Sendén, Sikström, and Lindholm 2015, Newman et al. 2008, Lenard 2016).

With the help of the Part of Speech Tagger implementation from the Python NLTK library, a list of pronouns found in all speeches was extracted (*NLTK :: nltk.tag package* 2021). The goal was to construct a list of female and male pronouns, as well as pronouns referencing the speaker or the audience. Table 4 shows the pronouns found throughout all speeches.

Female pronouns	Male pronouns	Pronouns without Gender	Not pronouns
she, her, hers, herself	he, him, his, himself	it, its, itself, me, myself, our, ourselves, their, theirs, them, themselves, they, us, we, you, you, yourself, yourselves	hope, mom, now, out, yes

Table 4: Pronouns found by NLTK's Part of Speech Tagger.

However, not all words classified as pronouns are actually pronouns, such as the word "now". Based on the list of pronouns produced with the help of NLTK's Part of Speech Tagger, I constructed a list of pronouns. I eliminated such words and added gendered words that are not pronouns to create lists of gendered and



```
Release eleven of June11942 Commencement Address of Governor
Julius P Veil University of Wisconsin Commencement June 1 1942
Or President Members of the Board of Events Members of the
Radiating Glass Parents and Friends It commencement time a year
ago the skies were overcast It had become evident to some that a
mighty storm was gathering Commencement Day1942finds America in
the midst of the most titanic struggle the world has ever known
For more than a year before the outbreak of war this state made
preparations for what might come Both by legislation and by
organization this commonwealth made ready for the conflict His
university too has done its part It has offered its facilities
to the armed forces of the United States and has sent many of
its sons to the line of conflict Today you are honoring a great
American You are conferring upon Douglas MacArthur the degree of
Doctor of Laws A great soldier General George Grunert is
receiving the degree upon behalf of Douglas MacArthur Wisconsin
knows MacArthur For years he lived among us He lives today in
the soul and spirit of every American He trifles to us the ideal
of American manhood Courageous intelligent resourceful
symbolized America in her resolute determination You who are
granulating have been the beneficiaries of free inst nations You
have lived in  land where there has been no limitation upon your
chances to better your position in life You have enjoyed the
Freedom to follow your conscience S dictator the right to be
yourself The ultimate value of this liberty depends upon the use
to which it is put It can be the means of moral and spiritual
advancement or the avenue of folly and decay It has been said
```

```
Release eleven AM June11942 Commencement Address of Governor
Julius i Heil University of Wisconsin Commencement June 1 1942 my
President Members of the Board of Regents Members of the
Graduating Class Parents and Friends At commencement time a year
ago the skies were overcast It had become evident to some that a
mighty storm was gathering Commencement Day1942finds America in
the midst of the most titanic struggle the world has ever known
For more than a year before the outbreak of war this state made
preparations for what might come Both by legislation and by
organization this commonwealth made ready for the conflict This
university too has done its part It has offered its facilities to
the armed forces of the United States and has sent many of its
sons to the line of conflict Today you are honoring a great
american You are conferring upon Douglas MacArthur the degree of
Doctor of Laws A great soldier General George grunt is receiving
the degree upon behalf of Douglas MacArthur Wisconsin knows
MacArthur For years he lived among us He lives today in the soul
and spirit of every American He typifies to us the ideal of
American manhood Courageous intelligent resourceful symbolizes
America in her resolute determination You who are graduating have
been the beneficiaries of free instil tuitions You have lived in
land where there has been no limitation upon your chances to
better your position in life You have enjoyed the freedom to
follow your conscience i dictator the right to be yourself The
ultimate value of this liberty depends upon the use to which it
is put It can be the means of moral and spiritual advancement or
the avenue of folly and decay It has been said that in a from
```

Figure 6: Comparison of the spell corrected version of a University of Wisconsin speech. On the left is the spell checked version using the Python spellchecker module and on the right the version produced with the Textblob spell corrector.

neutral features. Using the words determined to be male and female pronouns, I constructed my own list of female and male features as seen in Table 5.

Female features	**Male features**	**Features that are not gendered**
she, her, she, her, hers, herself, woman, women, female, females, girl, girls	he, him, his, himself, man, men, male, males, boy, and boys	it, its, itself, me, myself, our, ourselves, their, theirs, them, themselves, they, us, we, you, you, yourself, yourselves

Table 5: Gendered pronouns used to extract female and male features from a speech.

In order to create a measure of how often gendered features occurred, the ratio of a particular type of feature and all features in Table 5 appearing in each speech is calculated. For example, in the sentence "**She** was the biggest influence in **my** life and helped **me** decide what career path **I** wanted to take", there are a total of four features that appear in Table 5. Of these features, one falls into the female features category, and thus the measure for female features is 1/4. In the case of male features, the score is 0 since there are no male features in that particular sentence.

To compare the measure of female to male features, both the ratio of female to male features and log-ratio was considered a measure of comparison. However, for some speeches, this ratio had a zero denominator. Therefore, the difference of female and male features was chosen as a final measure to gauge if a speech had more male or female references.

In addition to focusing on male and female features, pronouns that mention the speaker versus the audience were considered. Again, words were inspired by the extracted pronouns with the help of the NLTK Part of Speech Tagger and pronouns seen in Table 4. Table 6 shows the pronouns that fall under each category.

Personal pronouns	**Audience pronouns**
I, my, me, myself	you, yours, yourself, and your

Table 6: Speakers versus Audience features.

The goal behind considering both personal pronouns and words referring to the audience was to gauge how interactive the speaker was throughout their speech and if a speaker focused more on their personal story or on the group they were addressing.

As with the gendered features, a ratio was calculated for the audience and personal features by taking the total number of audience and personal pronouns mentioned in a text and comparing it to the count of either audience or personal pronouns. For example, in the sentence "If all this is true, it is clear, **I** think, that **you** are going to have to do hard things with **your** minds in the next twenty years", there are a total of three relevant pronouns of which one falls into the personal category. Thus the ratio, in this case, would be 1/3.

It is important to note that this list of considered pronouns is not comprehensive. Some pronouns can be ambiguous. For example, the pronoun "they" can refer to a group of individuals or a single person. In addition, there may be other pronouns that were not included in the list that would have made it more complete.

6.2 Sentiment Analysis

A commencement speech is usually associated with a positive message. Many qualitative analyses of commencement speeches generally show that commencement speeches try to evoke positive emotions (Partch and Kinnier 2011). To gauge the sentiment of a speech, tools including the SentimentIntensityAnalyzer package from the Python NLTK natural language processing library were used (*NLTK :: nltk.tag package* (2021)). With the analyzer's help, each speech received an intensity score for positive, neutral, and negative sentiment on a continuous scale between 0 and 1. The classification is based on the general use of words classified with a happy or positive sentiment versus a more sad and negative feeling. Different sentiments may be evoked across the progression of the speech. The overall sentiment of a speech might thus not reflect the change in sentiment throughout any particular speech. To examine the sentiment of a speech as each speech progressed, each speech was broken down into ten equal parts of equal word length. Each part was then assigned a positive, neutral, and negative sentiment score, parallel to the sentiment classification of the whole speech. The choice of dividing a speech into ten parts was made after trying the different number of parts. It was found that five parts did not offer enough fine-grained information about the sentiment. On the other hand, fifteen parts led to some parts of some speeches having only a few sentences, thus capturing too many small fluctuations across small time intervals.

6.3 Evaluating Topics

Bigrams were extracted for all of the cleaned text by counting the number of times adjacent word pairs occurred. From all bigrams, the 50 bigrams that occurred the most overall speeches were chosen for further analysis. These bigrams were manually grouped into ten categories as seen in table 7. A subset of bigrams was put into a miscellaneous category. They included words such as archival, digitized, archive, which were likely remnants of the transcription of headers and footers of the speeches. All words were lowercased and stemmed before calculating the bigram frequencies. Therefore certain words like unite in the topic "United States" appears in this form.

The idea behind grouping different bigrams was to model the topic creating an algorithm that other methods such as LDA use. For each speech, the number of times a specific bigram and topic is mentioned was counted and normalized by taking the speech length into account (dividing by speech length). Therefore, the ratio of how much a particular topic is mentioned for each speech could be measured.

7 Results for All Colleges

In this section, the research questions introduced Section 1 will be addressed. Section 7.1 will answer question 1, Section 7.2 examines question 2, questions 5 and 6 will be examined within each of the subsections.

Figure 7 shows the speakers and their genders at each college in each year. From Figure 7, one can see that starting around roughly the 1970s, the number of female speakers increased and shifted to include solely female speakers at most women's institutions. Female speakers appeared only at women's colleges until 1967, when Barbara Ward Jackson was the commencement speaker at St. John's University, a men's college.

Topic	Bigrams
People	man woman, woman man, young woman, young people, family friend, young man
International Politics	united nation, world war
Education	high education, liberal art, board trustee, alma mater, law school, college university, high school
United States	united state, white house
Civil Action	civil right, human right, human being, public service, change world, real world, role model
Future Advice	work hard, hard work, ask question, live life, find way, point view
Graduation	college graduate, commencement speaker, commencement speech, commencement address
Temporal	past year, year old, year ago, year later, thirty year
Healthcare	health care
Miscellaneous	bryn mawr, pomona college, mount holyoke, new hampshire, president brazil, cwru archive, document digitize, archival original, new york, woman college, take place, look like

Table 7: Grouped topics based on bigram extraction approach.

7.1 Speech Lengths

The mean length of the 825 speeches is 2,710 words. Looking at the average speech length at each college showed some differences in length. This patterns likely occurs because each college provided speeches for different years. In Figure 8, one can see that the earlier speeches at Stanford are very long in comparison with other speeches, showing the average speech length can vary by college and the years for which speeches exist. The overall speech length has varied over time, but in general, a trend of speeches becoming shorter over time can be discerned, as seen in Figure 8.

7.2 Use of Pronouns

Research question 3 focuses on pronoun usage. There were a few colleges with more female than male features, which were Brown, Barnard, College of St. Benedict, Mills College, Mount Holyoke, Scripps College, Smith College (the one with the most), and Wellesley College. At Smith, the difference between female and male pronoun percentage was 0.1387 . Out of this group, all colleges except Brown are women's colleges. All college types except the seven sister colleges used more male than female features. Especially the conservative colleges had a higher rate of male features to female features than other colleges. Over the decades, one can see a shift to using more female features. Figure 9 visualizes this trend over time.

In addition, the larger the institution, the more male features were included in speeches (ratio of 0.08 more male than female features). However, this is most likely due to the fact that most of the smaller liberal arts colleges also included all of the female speakers and all women's institutions. When the speeches were grouped by gender, one could see that female speakers used more female pronouns in their speeches than male speakers over time (see Figure 11). However, grouping the speeches by college type did not show any visible trends.

The audience to personal feature measure does not follow an overall pattern across time. However, if the speeches are grouped by college type, the speeches from the HBCU and Ivy League colleges seem to have more audience rather than personal features (see Figure 10).

For the personal to audience features, the audience was addressed more in large (value of 0.03 more audience to personal features) and medium-sized universities versus smaller, where people discussed themselves more (value of 0.01 more personal to audience features).

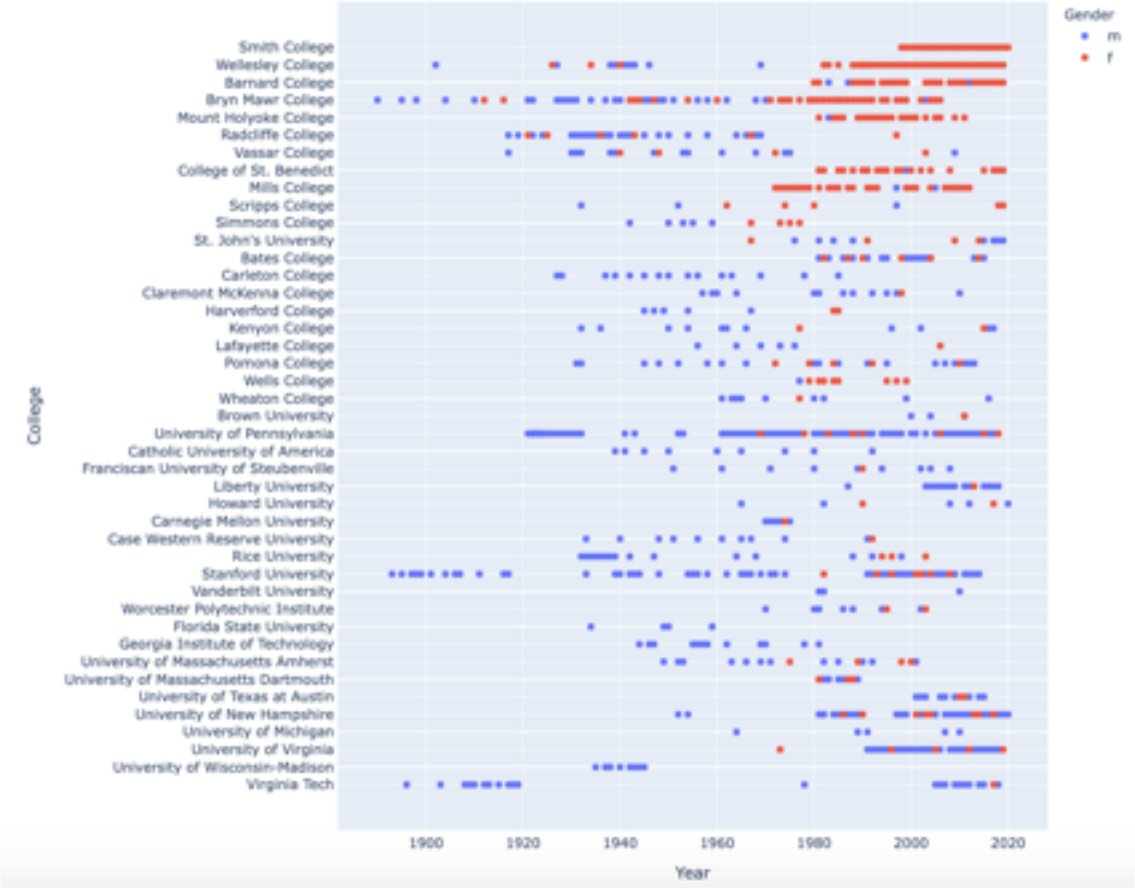


Figure 7: Speeches over the years by colleges.

7.3 Speech Sentiment

Research question 3 on how sentiment varied over time and different institutions. Overall, the speech sentiment did not vary substantially between different types of colleges. As in the case study for Wellesley College, the average sentiment was positive overall. However, the average sentiment was more negative years during the 1970s and during the Second World War (1930s and 1940s). Figure 13 displays this trend.

Aside from looking at the general speech sentiment, the speech sentiment by the gender of the speaker and by college type can be examined. For example, Figure 15 shows that, on average, the sentiment is slightly more positive in the speeches given by male speakers. There appears to be a stark spike in positive sentiment in the 1930s in the speeches given by female speakers, but this happens because there were only two female speakers observed in that decade. In addition, Figure 16 shows how the sentiment of speeches changed based on college type.

Looking at the individual speeches and the parts of the speech, Figure 14 shows how the positive, negative and neutral sentiment changes on average over the progression of a speech. Overall, the average sentiment does not very much over the progression of a speech. However, there is a slightly more positive trend at the beginning and the end of a speech, similar to the Wellesley case study.

7.4 Topics of Speeches

Research question 4 focuses on the topics mentioned in commencement speeches. With the help of the bigram approach detailed in Section 6 and listed in Table 7, the most salient topics mentioned in the speeches were

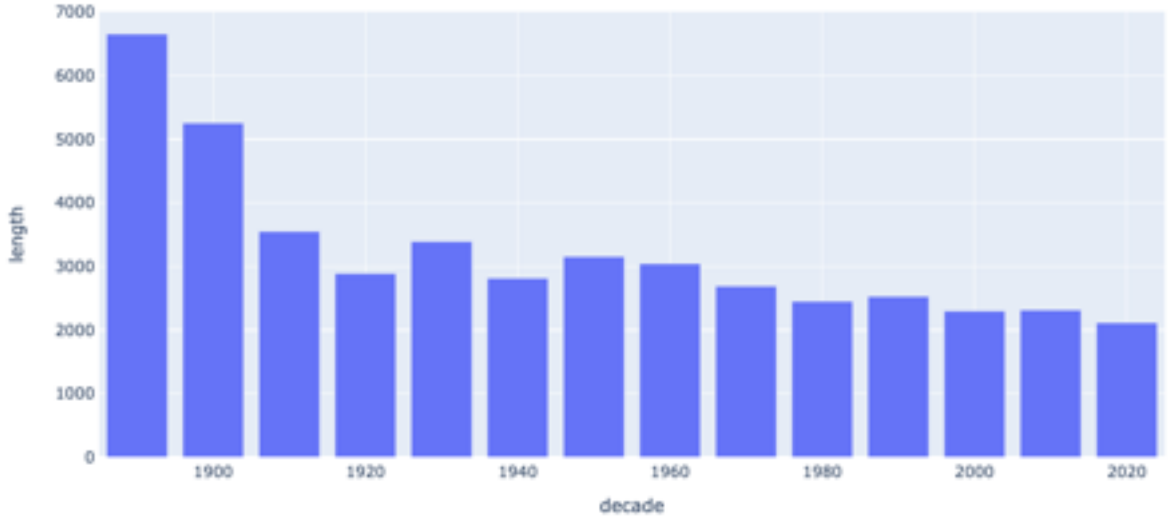


Figure 8: Average speech length by decade

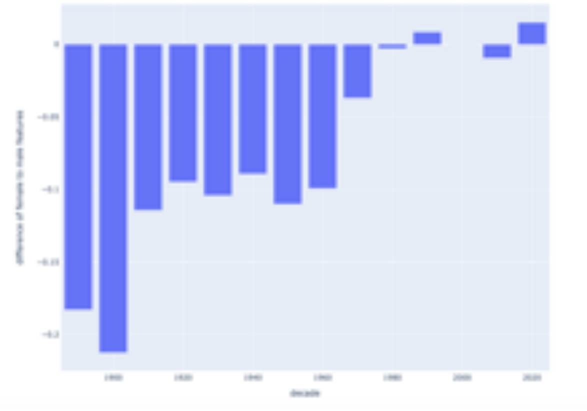


Figure 9: Speeches over the years and by difference between female and male features.

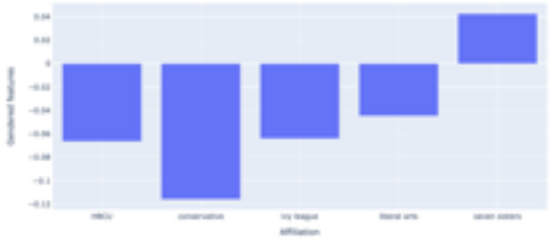


Figure 10: Average gendered features by college type.

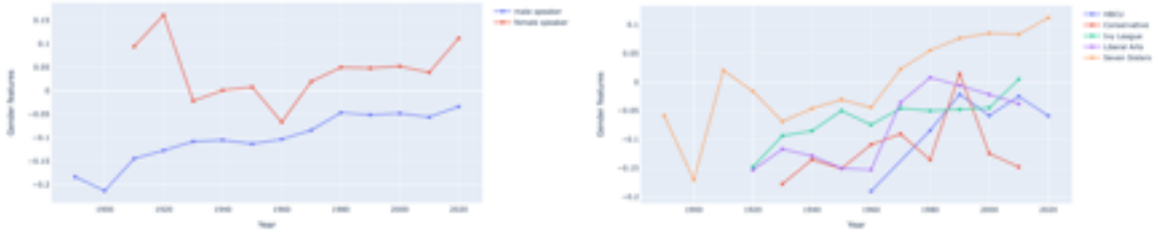


Figure 11: Difference between female and male features over time, by the speaker's gender and type of college.

determined. I calculated the correlation between the topics and other measures such as the length of the speech, sentiment, and pronouns. The topic "United States" had the highest correlation with topic of foreign institutions (0.27) and a correlation of 0.17 with the topic of people, positive sentiment (0.165), and civil rights (0.136), and difference in sentiment (0.127). In Table 8, the correlations between the different topics

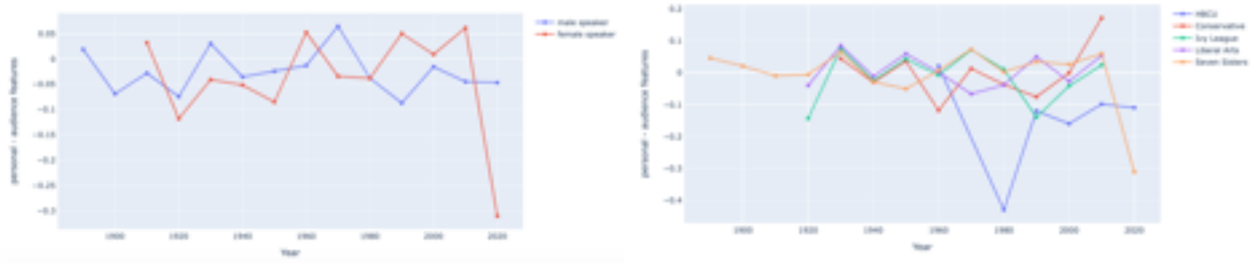


Figure 12: Use of personal-to-audience features over the years by speaker and college type.

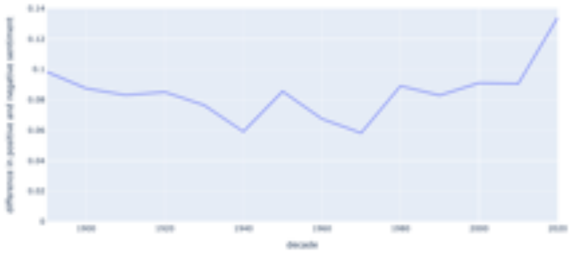


Figure 13: Speeches by the difference between positive and negative sentiment over the decades.

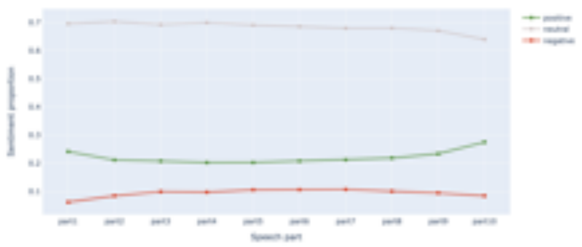


Figure 14: How the presence of positive, negative and neutral sentiment varies over the progression of a speech.

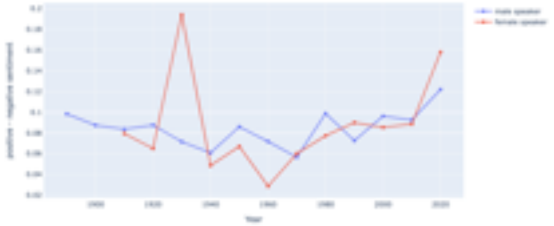


Figure 15: Speeches by the gender of the speaker.

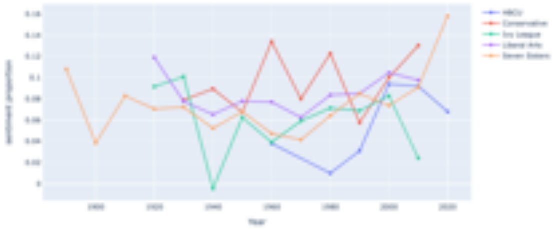


Figure 16: Speeches by college affiliation and sentiment over time.

and other topics can be seen.

	year	female features	male features	personal features	audience features	diff. gender	decade	positive	negative	diff. pers.-aud.	diff. sentiment	man woman	united nation	united state	civil right
man woman	-0.003	0.172	0.083	-0.035	0.114	0.056	0.002	0.165	0.022	-0.114	0.127	1.000	0.004	-0.032	0.137
united nation	0.001	-0.013	-0.027	-0.048	-0.074	0.010	-0.003	0.000	0.158	0.022	-0.093	0.004	1.000	0.273	0.102
high education	0.043	0.013	-0.032	0.062	0.077	0.031	0.042	0.121	-0.191	-0.015	0.216	0.065	-0.078	0.031	0.005
united state	0.035	-0.046	-0.014	-0.041	-0.129	-0.021	0.043	0.106	0.102	0.070	0.030	-0.032	0.273	1.000	0.060
civil right	0.210	0.077	-0.127	0.034	0.089	0.137	0.216	0.072	0.077	-0.044	0.016	0.137	0.102	0.060	1.000
work hard	0.192	-0.017	-0.110	0.092	0.235	0.064	0.196	0.120	-0.015	-0.115	0.111	0.046	-0.024	-0.061	0.017
health care	0.134	0.007	-0.010	0.070	0.055	0.073	0.133	0.033	0.012	0.008	0.021	-0.036	-0.017	0.051	0.032
college graduate	-0.013	-0.017	-0.046	-0.025	-0.029	0.022	-0.017	-0.106	-0.078	0.004	-0.043	-0.013	0.006	-0.018	-0.013
past year	0.274	0.033	-0.044	0.270	0.161	0.052	0.273	0.135	-0.009	0.072	0.120	0.059	-0.001	0.089	0.065

Table 8: Correlations of different topics and other examined values.

Figure 17 shows how the topic of civil rights was mentioned in the various types of colleges over time. Although these correlations are small, they as still worth considering. Interestingly, in the speeches of the conservative colleges, civil rights was mentioned relatively frequently in the 1970s. However, due to the small number of colleges falling under this group, it is most likely due to the small sample size. In general, though the correlation between a topic and a sentiment was measured it does not necessarily imply that the speaker was speaking about that topic with that sentiment.

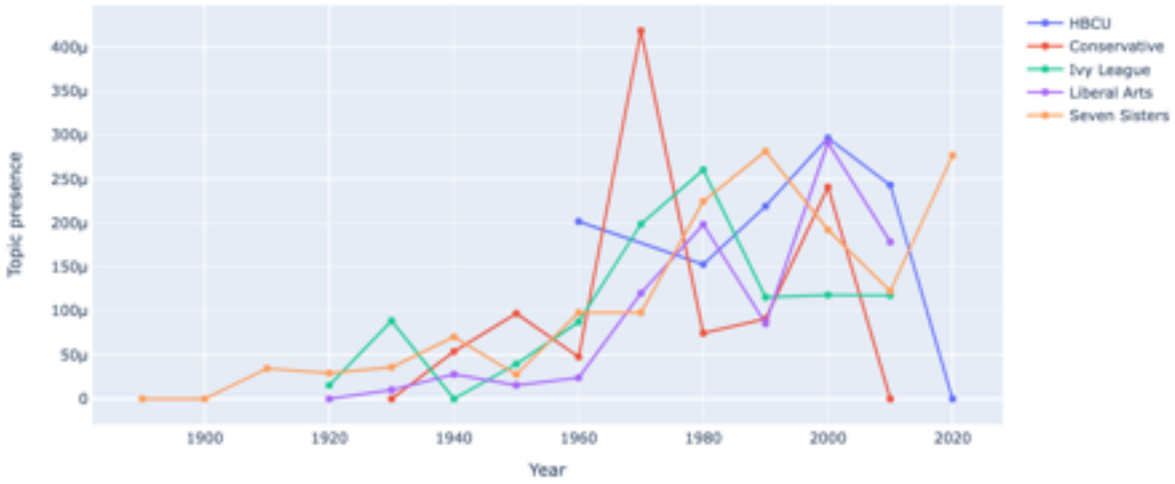


Figure 17: How often the topic of civil rights is mentioned by college type.

8 Discussion, Limitations and Future Work

8.1 Discussion

This paper was motivated by the idea that commencement speeches provide a unique source of information about what commencement speakers share with young adults in a transitional period of their lives. This analysis shows the change in time in terms of speech lengths and a difference in the use of gendered features based on the gender of a speaker, where female speakers used more female features than male speakers. This suggests that female speakers might have been more aware of gender in their speeches or mentioned influential women in their lives. In general, female speakers appeared more frequently at women's colleges and thus might have also addressed a more female audience.

Although the topic analysis mainly focused on bigrams, the extracted topics indicated a relationship between different features that were not apparent at first sight. For example, the topic of civil and human rights had a correlation of 0.216 with the decade and a correlation of 0.136 with the variable gender features. Although small, these correlations could indicate that the topic of "civil rights" was more important in certain years and directly related to the gender features a speaker used in their speech.

8.2 Limitations

There are areas of the data collection, data processing, and analysis process that warrant further consideration and have possible limitations.

8.2.1 Data Collection

The first limitation involves the data that was collected. Although an attempt was made to include speeches from a variety of colleges that fall into different categories, the number of colleges was small for some types. Thus, the collected dataset was limited mainly by speeches made available via college archives or websites with commencement speeches. In addition, the data collection process was limited by the COVID pandemic as many college archives were closed at the time of data collection and thus were unable to provide the data for their colleges. Also, the colleges that provided speeches, including Wellesley, did not have speeches for every year or every decade. Therefore, the dataset is not representative of either colleges or years.

8.2.2 Data Cleaning and Processing

As described in Section 5, the data processing was complex. In addition, despite having multiple processing steps, there is still some messiness in the data, as seen in the example of Figure 3. As a result, some of the data were not useful, and tools like AWS Textract were unsuccessful in transcribing some of the speeches. In addition, other text transcription software such as the Google Vision API did not perform with a higher transcription accuracy. Due to this issue, all of the speeches from Dartmouth College had to be removed from consideration.

8.2.3 Analysis

Using methods such as PCA and BERT did not work as expected, although they have been proven helpful in other text analysis contexts. Therefore, this paper's approach to extracting topics relied primarily on bigram frequencies, but other topic modeling methods could lead to additional insights.

8.3 Future Work

For this paper, a large dataset of commencement speeches was collected, processed and analyzed. All three of these steps could be expanded in future work. Collecting a more extensive dataset could yield better and more generalizable results as some machine learning techniques require larger datasets for training. As described in Section 8.2.2, one of the issues was the transcription process. However, Berg-Kirkpatrick, Durrett, and Klein (2013) provide a way to create a generative probabilistic model that can learn various font structures and thus decipher and better transcribe historical text. Their study showed a 31% relative reduction in word error rate for commercial transcription software and showed a promising approach to better process the data used in this paper. Analyzing how the gender of the speaker and the audience combined with other measures could be a possible next step. Techniques including PCA and using a supervised version of BERT were tried on the dataset but failed to generate useful insights. Further exploring how these methods could be leveraged in a meaningful way in commencement speech analysis could allow for better topic classification.

9 References

[1] Yasamin Abbaszadegan. *NLP-Commencement-speech-Analysis.* original-date: 2020-01-07T17:21:17Z. July 2021. URL: https://github.com/YasaminAbbaszadegan/NLP-Commencement-speech-Analysis (visited on 09/13/2021).

[2] Rihab Abduljaleel Saeed Alattar. “A Speech Act Analysis of American Presidential Speeches”. In: *Arts Journal* 110 (2014), pp. 1–39.

[3] Bhagyashree Vyankatrao Barde and Anant Madhavrao Bainwad. “An overview of topic modeling methods and tools”. In: *2017 International Conference on Intelligent Computing and Control Systems (ICICCS).* June 2017, pp. 745–750. DOI: 10.1109/ICCONS.2017.8250563.

[4] Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein. “Unsupervised transcription of historical documents”. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* 2013, pp. 207–217.

[5] Justin Berkman. *The 70 Most Conservative Colleges in America.* URL: https://blog.prepscholar.com/most-conservative-colleges (visited on 02/27/2021).

[6] Ewa Bogdanowska-Jakubowska. “The discursive construction of high achievers’ identities in American culture”. In: *Lodz Papers in Pragmatics* 14.2 (2018). Publisher: De Gruyter, pp. 249–271.

[7] Ewa Bogdanowska-Jakubowska. “The discursive representation of places significant for an individual: an analysis of Polish academic year inauguration speeches and American commencement addresses”. In: (2021). Publisher: Katowice: Wydawnictwo Uniwersytetu Ślaskiego.

[8] Venkateswarlu Bonta and Nandhini Kumaresh2and N. Janardhan. “A comprehensive study on lexicon based approaches for sentiment analysis”. In: *Asian Journal of Computer Science and Technology* 8.S2 (2019), pp. 1–6.

[9] Suzanne Bordelon. “Composing women’s civic identities during the progressive era: College commencement addresses as overlooked rhetorical sites”. In: *College Composition and Communication* (2010). Publisher: JSTOR, pp. 510–533.

[10] *Commencement Archives — Wellesley College.* URL: https://www.wellesley.edu/events/commencement/archives (visited on 02/08/2021).

[11] *Digest of Education Statistics, 2016.* EN. Publisher: National Center for Education Statistics. URL: https://nces.ed.gov/programs/digest/d16/tables/dt16_303.90.asp.

[12] *Find a College Search — Women’s College Coalition.* URL: https://www.womenscolleges.org/colleges/default.htm (visited on 07/01/2021).

[13] Pollyanna Gonçalves et al. “Comparing and combining sentiment analysis methods”. In: *Proceedings of the first ACM conference on Online social networks.* 2013, pp. 27–38.

[14] Katie Huang. *Beginner’s guide to an NLP project: Analysis of commencement addresses in the U.S.* en. Apr. 2021. URL: https://towardsdatascience.com/beginners-guide-to-an-nlp-project-analysis-of-commencement-addresses-in-the-u-s-5bf228c3c5e7 (visited on 11/15/2021).

[15] Clayton Hutto and Eric Gilbert. “Vader: A parsimonious rule-based model for sentiment analysis of social media text”. In: *Proceedings of the International AAAI Conference on Web and Social Media.* Vol. 8. Issue: 1. 2014.

[16] Jennifer C. Glover Konfrst. “Messaging strategies in presidential commencement speeches 1980-2016: A content analysis”. In: *Teaching Journalism & Mass Communication* 7.2 (2017). Publisher: Association for Education in Journalism and Mass Communication Small . . ., pp. 49–56.

[17] Dragana Bozic Lenard. “Gender differences in the personal pronouns usage on the corpus of congressional speeches”. In: *Journal of Research Design and Statistics in Linguistics and Communication Science* 3.2 (2016), pp. 161–188.

[18] *Length of inaugural addresses of U.S. Presidents 1789-2021.* en. URL: https://www.statista.com/statistics/243686/length-of-inaugural-addresses-of-us-presidents/ (visited on 04/19/2021).

[19] Bing Liu. “Sentiment analysis and opinion mining”. In: *Synthesis lectures on human language technologies* 5.1 (2012). Publisher: Morgan & Claypool Publishers, pp. 1–167.

[20] Steven Loria. “textblob Documentation”. In: *Release 0.15* 2 (2018), p. 269.

[21] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. “Sentiment analysis algorithms and applications: A survey”. In: *Ain Shams engineering journal* 5.4 (2014). Publisher: Elsevier, pp. 1093–1113.

[22] Rushab Munot and Ani Nenkova. “Emotion impacts speech recognition performance”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. 2019, pp. 16–21.

[23] *National Center for Education Statistics*. EN. Publisher: National Center for Education Statistics. URL: https://nces.ed.gov/fastfacts/display.asp?id=84 (visited on 06/30/2021).

[24] *National Center for Education Statistics*. EN. Publisher: National Center for Education Statistics. URL: https://nces.ed.gov/fastfacts/display.asp?id=667 (visited on 06/30/2021).

[25] Matthew L. Newman et al. “Gender differences in language use: An analysis of 14,000 text samples”. In: *Discourse processes* 45.3 (2008). Publisher: Taylor & Francis, pp. 211–236.

[26] *NLTK :: nltk.tag package*. URL: https://www.nltk.org/api/nltk.tag.html (visited on 03/21/2021).

[27] Michael Nokel and Natalia Loukachevitch. “A method of accounting bigrams in topic models”. In: *Proceedings of the 11th workshop on multiword expressions*. 2015, pp. 1–9.

[28] Peter Norvig. *How to Write a Spelling Corrector*. URL: https://norvig.com/spell-correct.html (visited on 06/14/2021).

[29] Martha Maria Papadopoulou, Anna Zaretskaya, and Ruslan Mitkov. “Benchmarking ASR Systems Based on Post-Editing Effort and Error Analysis”. In: *TRITON 2021* (2021), p. 199.

[30] Jenifer J. Partch and Richard T. Kinnier. “Values and messages conveyed in college commencement speeches”. In: *Current Psychology* 30.1 (2011). Publisher: Springer, pp. 81–92.

[31] Sergio Peignier and Patricia Zapata. “Analysis of Fidel Castro Speeches Enhanced by Data Mining”. In: *Digital Humanities Benelux Journal* (2019).

[32] Nicole Peinelt, Dong Nguyen, and Maria Liakata. “tBERT: Topic models and BERT joining forces for semantic similarity detection”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 7047–7055.

[33] Prabhat Rayapati. *rooster06/UVA-CommencementSpeeches: BiGram language model and Readability Analysis*. URL: https://github.com/rooster06/UVA-CommencementSpeeches (visited on 03/09/2021).

[34] Filipe N. Ribeiro et al. “Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods”. In: *EPJ Data Science* 5.1 (2016). Publisher: Springer, pp. 1–29.

[35] Markella B. Rutherford. “Authority, autonomy, and ambivalence: moral choice in twentieth-century commencement speeches”. In: *Sociological Forum*. Vol. 19. Issue: 4. Springer, 2004, pp. 583–609.

[36] Marie Gustafsson Sendén, Sverker Sikström, and Torun Lindholm. ““She” and “He” in news media messages: Pronoun use reflects gender biases in semantic contexts”. In: *Sex Roles* 72.1-2 (2015). Publisher: Springer, pp. 40–49.

[37] *The Best National Universities in America*. en. URL: https://www.usnews.com/best-colleges/rankings/national-universities (visited on 02/21/2021).

[38] Ethan C. Tucker, Colton J. Capps, and Lior Shamir. “A data science approach to 138 years of congressional speeches”. In: *Heliyon* 6.8 (2020). Publisher: Elsevier, e04417.

[39] William Ughetta and Brian W. Kernighan. “The Old Bailey and OCR: Benchmarking AWS, Azure, and GCP with 180,000 Page Images”. In: *Proceedings of the ACM Symposium on Document Engineering 2020*. 2020, pp. 1–4.

[40] Julien Velcin, Mathieu Roche, and Pascal Poncelet. “Shallow text clustering does not mean weak topics: How topic identification can leverage bigram features”. In: *DMNLP: Data Mining and Natural Language Processing*. Vol. 1646. 2016.

[41] Yuting Zhu. “An Intercultural Analysis of Personal Metadiscourse in English and Chinese Commencement Speeches.” In: *Advances in Language and Literary Studies* 9.5 (2018). Publisher: ERIC, pp. 100–110.

10 Appendix

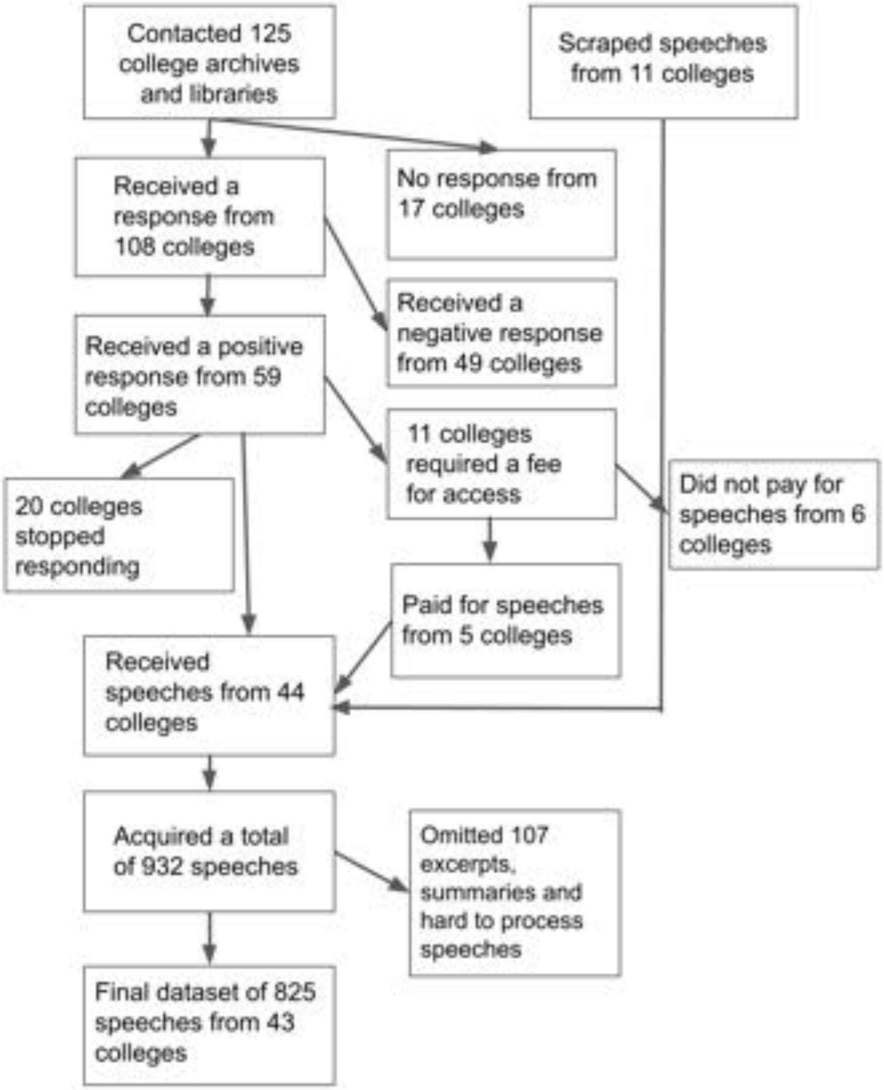

Contacted 125 college archives and libraries
Scraped speeches from 11 colleges
Received a response from 108 colleges
No response from 17 colleges
Received a negative response from 49 colleges
Received a positive response from 59 colleges
11 colleges required a fee for access
Did not pay for speeches from 6 colleges
20 colleges stopped responding
Paid for speeches from 5 colleges
Received speeches from 44 colleges
Acquired a total of 932 speeches
Omitted 107 excerpts, summaries and hard to process speeches
Final dataset of 825 speeches from 43 colleges

Figure 18: Description of data collection process.