Using Text-Mining Methods to Understand Guest Speaker Presentations in Statistics 141SL

Nikhil Sharma, Tiffany Trinh, Huiyu Chuang, Eddie Liu, Nicholas Ortega

Abstract

In the most recent quarter of Statistics 141SL, the professor had speakers with careers in statistics present to students on how they use statistics on a day-to-day basis in a professional setting. With data from responses to speakers from the past winter quarter, this report aims to delve deep into explaining how we can ensure the booking of effective speakers for Statistics 141SL and how we can determine whether responses are well thought-out. We start off by performing sentiment analysis on the full responses using the VADER Sentiment Analysis tool. We then clean and process our text data, and explore each variable independently and in relation to other variables. We use hierarchical clustering on our responses to analyze the themes of each speaker. We find that Mr. Sweetnam and Dr. Anderson deliver more pointed presentations than Dr. Kricorian, and thus recommend that speakers be asked to have a topic in mind prior to presenting to yield more worthwhile presentations. We use logistic regression to predict whether a response is well-thought out and thorough or not. Our model yields an accuracy rate of 77.45% and a p-value of 4.858 x 10^{-6} when compared to a no-information model. Based on our model, we recommend that students be given a minimum word count and topics to write about when given this assignment.

Problem Statement

Throughout the Winter 2020 quarter, there have been three guest speakers who have delivered presentations to Statistics 141SL classes: Mr. Quinn Sweetnam, a Senior Analyst of Decision Sciences at NBCUniversal Media; Dr. Ariana Anderson, an Assistant Professor of Statistics at the David Geffen School of Medicine at UCLA; and Dr. Karin Kricorian, a Director of Management Science & Integration at Walt Disney Parks & Resorts. All three of them gave pointed speeches to statistics students aspiring for careers in statistics post-graduation.

After each presentation, students were asked to submit responses to Professor Esfandiari that contained a short paragraph encompassing what they learned from the presentations. We were given all of the responses from the three speakers compiled in a Word document, from which we were given the open-ended task of analyzing the text data to uncover insights.

We decided to formulate our problem with two guiding questions:

1. How do we ensure we are booking effective speakers?

2. How do we determine whether responses are well thought-out or not, and how can we ensure thoroughness in future responses?

Variables of the Study and How They Were Measured

Our original dataset consisted of only two variables: “Comment” and “Speaker”, with the former denoting the responses made by each student and the latter denoting which guest speaker the response was geared towards. In order to gain more information for our analysis and for our model, we decided to create new variables using sentiment analysis and tagging.

Our sentiment analysis involved using the VADER Sentiment Analysis tool in Python to assign each comment a numeric score by sentiment. VADER is a lexicon and rule-based sentiment analysis tool that scores comments based on a variety of factors including the perceived sentiments of words themselves, punctuation, negative words, modifiers (such as “very” and “a lot”), conjunctions and more. Therefore, we fed the raw, uncleaned sentences to the VADER sentiment analyzer to take advantage of these features. Since in essence each response could be considered a review of the given presentation, we decided that VADER would be an adequate tool to use since it has shown to work well on New York Times editorials and movie/product reviews.

The sentiment values returned by VADER include positive, negative, neutral, and compound scores with the columns named respectively after each sentiment. The compound score is a normalized, weighted combination of the positive, negative and neutral scores. The numeric nature of the scores in these four columns signify the probability of each comment matching a certain sentiment, with higher values signifying higher probabilities and lower values signifying lower probabilities. For example, a comment with a value of 0.216 under the “positive” column indicates the comment has a 21.6% probability of being associated with a positive sentiment.

We then split up the 204 responses and had each member of the group manually read ~40 responses each and manually tag the responses as proper, thorough responses or not. Our criteria for what was considered a thorough response was if the response required the student to pay attention to more than an isolated section of the talk; i.e., the response covered a moderate breadth of topics discussed in the presentation. By using this criteria, we hoped to eliminate subjective biases in our own assessments of each response. We tagged comments with a 1 if they met the criteria and a 0 if they did not, creating a target binary variable. There is a basis in the natural language processing world to use manually tagged words and corpuses for later statistical analysis. A great example is the VADER tool, whose lexicon with perceived sentiments of words was aggregated through human volunteers on Amazon Mechanical Turk. Thus, we feel that our process here has some merit in the world of text mining.

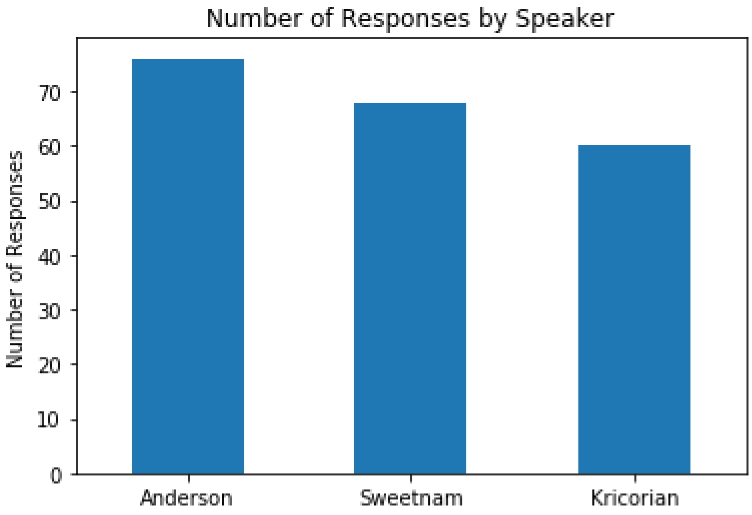
To create our other features, we cleaned the responses by removing stop words and grammar in addition to lemmatizing words. Stop words are words that are filtered out in natural language processing projects before analysis; examples include articles such as “the”, “a”, “an” etc. We extended our stop word dictionary (which exists in the *Natural Language Toolkit* module in Python) by adding words that repeatedly showed up in responses without adding anything relevant to analysis, such as “Mr. Sweetnam”, “Dr. Kricorian” and “Dr. Anderson”. Lemmatizing is the process of converting verbs to their unconjugated form for cleaner analysis (for example, changing the words “running” and “walked” simply to “run” and “walk”).

Finally, we used the cleaned words to come up with top unigram and bigram counts. We found the top 30 most common one-word and two-word phrases (unigrams and bigrams) per speaker, then counted the number of times those top 30 unigrams and bigrams showed up in a response given the appropriate speaker.

Exploratory Data Analysis

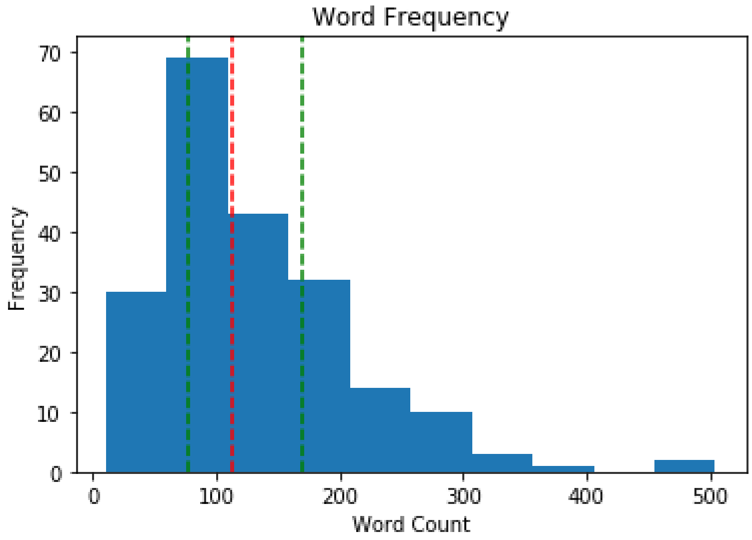
After processing and cleaning our data, we conducted some preliminary exploratory data analysis in order to better understand our data and the variables that could be used in our model. We started off by first implementing a correlation matrix to examine any potential relationships between all of our variables. We found that there was a strong negative correlation (-0.883) between neutral and positive scores, and there was also a strong positive correlation (0.869) between positive and compound scores. This implies that higher probabilities of neutral scores are generally associated with lower probabilities of positive scores, and higher probabilities for positive scores are generally associated with higher probabilities of compound scores.

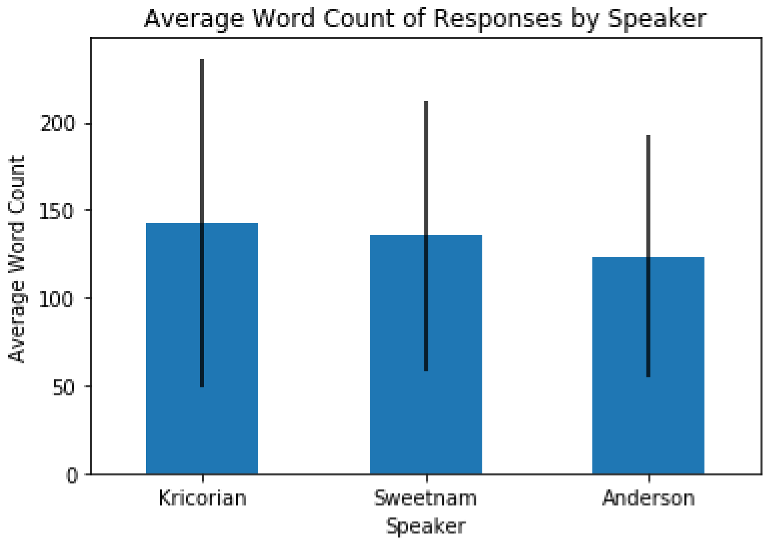
Moving on, we examined the number of responses by the speaker. Dr. Anderson had the most responses (76) followed by Mr. Sweetnam (68) and Dr. Kricorian (60) for a total of 204 responses.



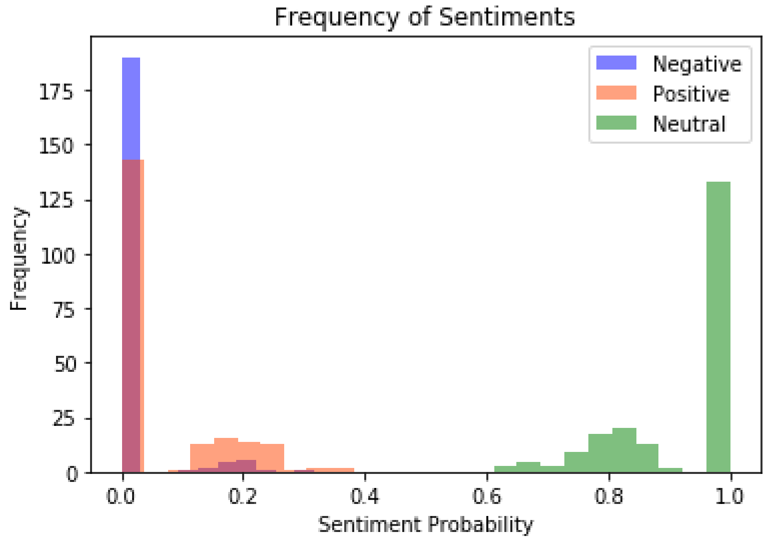
We also examined the distribution of the "Tag" variable. There are 128 (63%) responses tagged as good and 76 (37%) tagged as bad. Thus, a slight majority of the responses are categorized as good responses.

Next we looked at the distribution of the word frequency of the responses. The average number of words out of all of the responses was 112.5 Here we defined average as the median because the distribution is right-skewed. The red bar signifies the median of the distribution whereas the green bars signify the 25th and 75th percentiles. The second graph showcases the average word count of responses by speaker. Dr. Kricorian had an average of 142 words in her responses, Mr. Sweetnam had an average of 135 words in his responses, and Dr. Anderson had an average of 123 words in her responses. We include standard error bars which then indicate that there is much variation in the length of responses for each speaker.

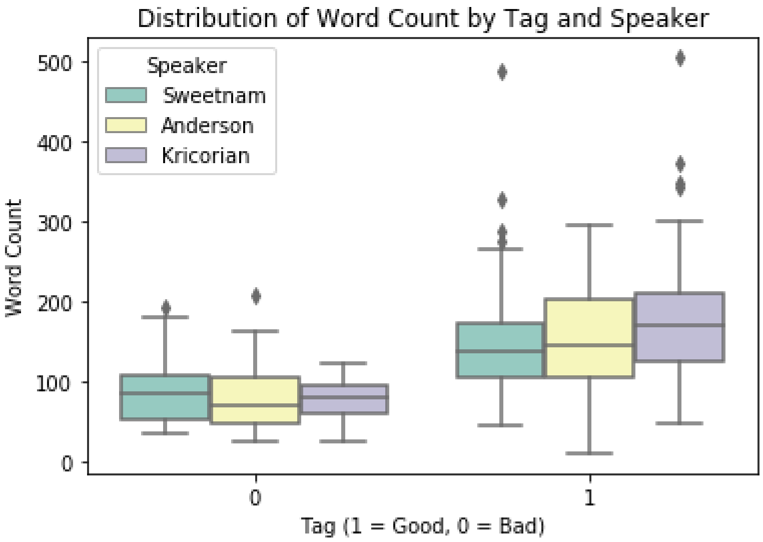




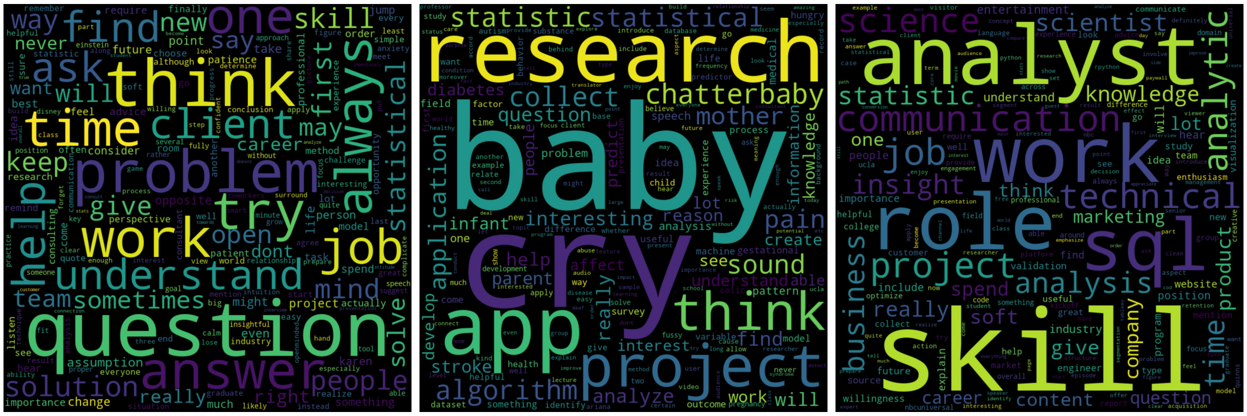
Next we examined the frequency of sentiment probabilities with respect to each sentiment. Most comments were labeled as purely neutral according to our sentiment analysis. Although the bars for positive and negative have the highest frequency of values, these all have probabilities of 0. The highest non-neutral sentiment frequency is that of Positive responses, centering around 20% probability (which signifies an 80% neutral probability). Very few comments are labeled as negative.



Our final graph showcases the distribution of word count in the responses, categorized by speaker and tag. We can see that regardless of Speaker, comments tagged as good generally have higher word counts. The word count by Speaker is generally very similar regardless of Tag.



As part of our exploratory data analysis, we created three word clouds to visual word usage for each speaker. From these word clouds we see that the speakers received fairly targeted comments. For example, Dr. Anderson's talk revolved around her work on the ChatterBaby app, which generally dealt with babies and crying. In Mr. Sweetnam's word cloud, we see that his words were also fairly targeted as students mentioned his emphasis on having a variety of skills such as SQL. Another seemingly memorable part of Mr. Sweetnam's talk was about the different analyst roles and positions along with the necessary skills for each role. At first glance, Dr. Kriocorian's word cloud seemed to feature words that were rather unspecific, however, the words do reveal that several topics were discussed. The word cloud for Dr. Kricorian showed that what students got out of the talk largely revolved around the importance of knowing and understanding the question before trying to solve the problem, which was emphasized by Einstein's quote that was prominent throughout the talk and course in general.



Left to right: Dr. Kricorian, Dr. Anderson, Mr. Sweetnam

We also found the most common bigrams from responses for each speaker. Here are the 10 most frequently used bigrams per speaker.

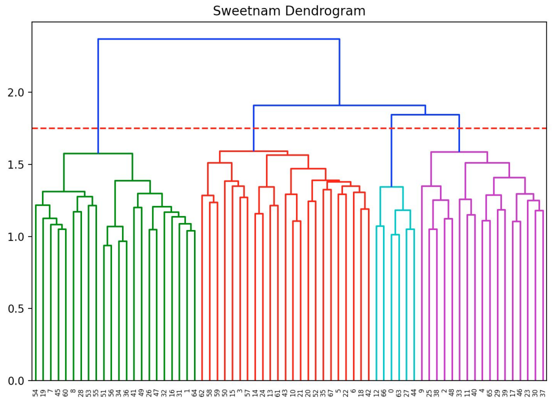
Table 1: Top 10 Most Frequently Used Bigrams by Speaker

Mr. Sweetnam	Dr. Anderson	Dr. Kricorian
technical skill	baby cry	open mind
soft skill	cry baby	solve problems
entertainment company	gestational diabetes	statistical consulting
time spend	research question	question ask
source validation	infant cry	new thing
communication skill	chatterbaby app	spend time
collect insight	cry sound	make sure
product content	statistical analysis	think question
validation analysis	substance abuse	understand question
communication enthusiasm	fussy hungry	question important

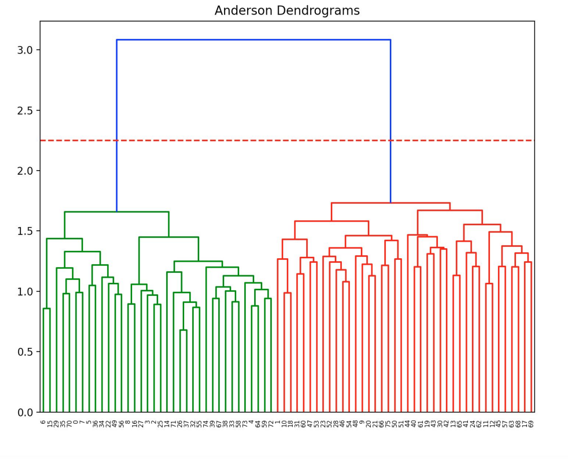
We again see how the most common phrases in each response guide the themes and content of each respective presentation.

Statistical Analysis: Hierarchical Clustering

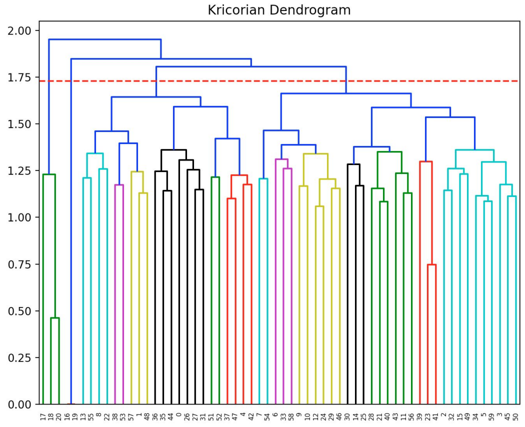
Our group used hierarchical clustering to further analyze the themes of each speaker. For each speaker, we tokenized all the words to collate an array of vocabulary words that are used throughout all the comments from that speaker. We then calculated the probability of each vocabulary word appearing in each comment. Using this probability data frame, we drew out three dendrograms - one for each speaker - to visualize the main themes for each speaker.



Looking at the dendrogram for Mr. Sweetnam, we see that there are 4 distinct clusters for this speech. Upon clustering all the comments and analyzing the different clusters, we see that each cluster has a distinct theme. The first cluster (green) of comments focused on post-grad career and jobs, the second cluster (red) focused on specific statistical skills, the third cluster (blue) focused on data science and data analysis and the last cluster (purple) focused on his experience in the entertainment field, specifically NBC Entertainment. There are some similarities between the third (blue) and fourth (purple) cluster, which include data analytics in the entertainment industry. These two clusters then link up with the second cluster (red) about skill sets, focusing on the technical skills in the workplace, specifically in the entertainment industry. When combining all four clusters together, we see how Mr. Sweetnam's talk focused on skill sets required for obtaining a post-grad data analytics role in the entertainment industry.



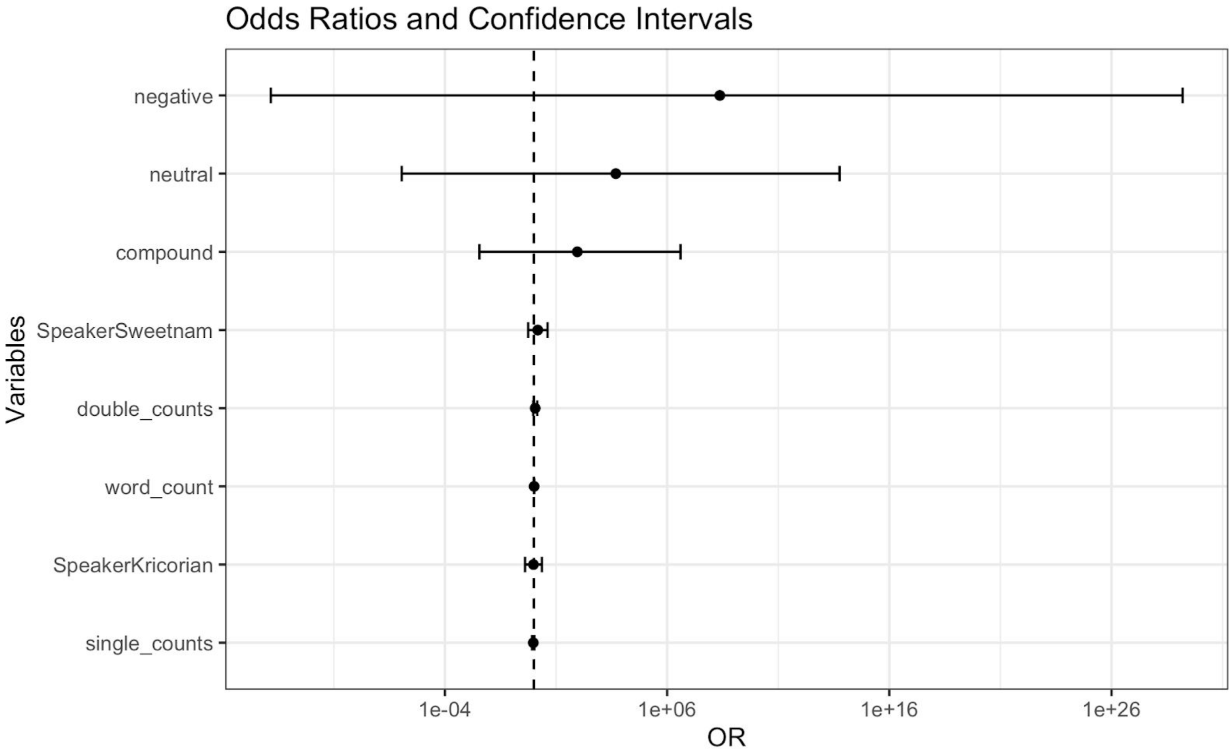
Looking at the dendrogram for Dr. Anderson, we see that there are only 2 distinct clusters for her speech. These two clusters focus mainly on Dr. Anderson's formal research in machine learning (green cluster) and her ChatterBaby app, involving babies and crying (red cluster). These two clusters show the overall theme of Dr. Anderson's talk, which focused on her research in machine learning and how it helped in developing the ChatterBaby app.



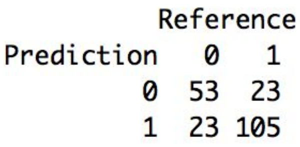
At an initial glance, it seems like Dr. Kricorian has many distinct clusters. This reflects what we see from the word cloud where there are not as many distinct words and themes. However, it seems like there are 4 main clusters. The first cluster focused on the idea of knowing the problem before finding the answer. The second cluster focused on soft skills at a workplace. The third cluster focused on the story about the fish and the water and the last cluster focused on being open to opportunities and not being afraid to say yes. Looking at the similarities in the clusters, the third and fourth clusters talk about being open to opportunity and accepting challenges. These two clusters are similar to the second cluster as they talk about the soft skills at a workplace, which include constantly seeking improvement, working with others, understanding people's work pace and being open to opportunities. Overall, the four clusters talk about how to develop the soft skills and the mindset required to excel as a statistical consultant.

Statistical Analysis: Logistic Regression

To analyze the quality of responses, we used logistic regression. We chose this type of regression to conduct our analysis because our outcome variable (Tag) was binary. We built a logistic regression model with Tag as our binary response variable and Speaker, Word Count, Negative Score, Neutral Score, Compound Score, Frequency of Top Unigrams and Frequency of Top Bigrams as our predictors. We excluded the Positive Score predictor because it created collinearity issues with the Compound Score predictor. Our model thus contained 1 categorical variable with 3 levels and 6 numerical variables.



Since our data only contained n = 204 observations, we decided to test our model using *k*-fold cross validation, with *k* = 10, rather than only testing on one holdout set. We hoped that by doing this, we would get less biased prediction results. Our model yielded an accuracy rate of 0.7745, meaning it misclassified observations about 22% of the time. The no-information rate (best guess given no information beyond the overall distribution of the classes you are trying to predict) was 0.6275 and the p-value was 4.858 x 10^{-6}. Our confusion matrix was as follows:



```
          Reference
Prediction   0   1
         0  53  23
         1  23 105
```

We can see here that our model outperforms simply picking the majority class for all test cases (the no-information rate). We can also see that with a p-value of 4.858 x 10^{-6}, the difference between our model and a no-information model is statistically significant.

We found relative variable importance of our model using the library *caret* in R, which ranks variable importance by ordering variables by the absolute value of their t-statistics. The results were as follows:

Table 2: Variable Importance of Logistic Regression Model

Variable	t-statistic, Absolute Value
Word Count	100.00
Frequency of Top Bigrams	20.18
Frequency of Top Unigrams	13.27
Compound Score	12.50
Negative Score	11.73
Speaker: Sweetnam	11.62
Neutral Score	10.64
Speaker: Kricorian	0.00

Overall Conclusions

Through our analysis, we saw three distinct themes across the three different speakers. Through these three distinct themes, we were able to analyze different aspects of these talks through word counts, word clouds, hierarchical analysis and logistic regression. All of these helped us explore the differences between the speakers, which included the number of topics each speaker touched on, the overall sentiments on each speaker and the focuses of each speaker. This also allowed us to analyze the comments for each speaker. Each speaker touched on many different topics, and tools such as hierarchical analysis helped us better understand the distinct and overlapping themes mentioned in the students' comments. We took a step further to better analyze the students' work throughout the quarter. Using factors like word count and n-gram frequency counts, we were able to predict whether a student put in effort in his/her response. We believe that our model did a formidable job on the predictions, given the statistically significant difference between our model and a no-information model.

Challenges of the Study

One of the main challenges we faced during our data analysis process was the general lack of variables that were included in our dataset. The data our team was provided with included only the comment itself and the name of the speaker it was written for. Retroactively requesting additional information from the commenters was not feasible, as the professor would have had to trace who submitted which comment and subsequently collect information from that student. This would have likely taken too much of the professor's time, so we proceeded without the information. However, having that additional data on the students who wrote each comment could have been beneficial in building our models. Stronger models would have helped in analyzing the data to gain a better understanding of what students learned from each speaker and how the professor can tailor speakers to fit student needs in the future. Basic information such as age, gender, post-graduation plans, and GPA would have aided in creating better predictive and more meaningful models. For example, if a student was not planning on going to graduate school, they may have found Mr. Sweetnam's talk more helpful, because the other two talks were given by PhDs in careers that reflected that level of education. Although there was not much we could do to gain more information about the commenter, we were able to create some of our own predictors to help improve our model.

Another challenge we faced involved the comments themselves. First of all, because the comments were freeform (due to the nature of the assignment), each one of them was formatted differently and required quite a bit of cleaning and removing of special characters and non-essential numbers. Each comment also needed to be read over (by the team) and tagged as either "good" or "bad". A "good" comment was defined by our standards as one that mentioned topics or specific parts of a talk that required the commenter to actually have attended the talk. A "bad" comment indicated that the comment was too general (not speaker specific) or too vague (could have been written even without personal attendance of the talk). This tagging was subjective but allowed us to create a better and more meaningful model.

The nature and complexity of text mining led to many difficulties in terms of working with the comments in our dataset. Our team read through all the provided comments to pick out irrelevant words to include in our list of stopwords, as it is important to have an accurate understanding of common words and phrases used for each respective speaker. Many of the most common stopwords (like "the", "and", and "a") were pre-generated and available in Python's *Natural Language Toolkit* module, but a handful were manually selected by the team (such as "thing", "talk", and "important"). We also had to implement a lemmatizer (a tool used for reduction of inflectional forms to stems or base forms) to ensure that variants of the same word would not be counted separately.

Recommendations for the Future

For future speaker bookings, we recommend that the professor have each speaker come in with a pointed presentation to ensure that the experience is meaningful for students. We saw in our clustering that the responses generally split nicely for Mr. Sweetnam and Dr. Anderson, but did not split as cleanly for Dr. Kricorian. For Dr. Kricorian, there were two clusters (clusters three and four) that seemed to overlap in terms of content, with both covering the ground of openness to opportunity and acceptance of challenges. This could indicate that the more free-flowing structure of Dr. Kricorian's presentation was not as valuable to students, since distinctions between topics were far less clear-cut than they were for Mr. Sweetnam and Dr. Anderson's presentations. Thus, to ensure that the experience is worthwhile for both presenters and students, it might pay to have the speakers come in with a specific topic or list of topics in mind.

For future response assignments similar to this one, we would recommend implementing a word count minimum. According to our model, word count was the most important predictor of whether the comment was tagged as "good" or "bad". All 59 comments in the validation sets that had a word count of at least at the average word count for "good" comments (163 words) were tagged as a "good" comment by the model. We acknowledge that our criteria for whether a response was thorough or not could be intrinsically tied to its length, as a higher word count implies that a responder could have discussed more topics. We attempted to avoid this confoundment by not automatically giving long responses a tag of 1, since many long responses drew on singular aspects of the presentations. Still, we understand that our tagging process might have been biased when it came to including word count as a predictor. In the future, we could potentially implement a more systematic, stringent criteria for manual tagging.

Another recommendation from our team for this assignment in the future would be to provide students with a list of topics that were discussed by the speakers in their talks. Ensuring that students' comment on a given list of topics would likely increase the number of "good" comments per speaker. From our bigram analysis and word clouds, we saw that for each speaker, the most prominent words were the overall topics within each talk. For example, the word cloud generated for Dr. Anderson's talk prominently displays the words "cry" and "baby" in accordance with Dr. Anderson's discussions of her work on the ChatterBaby app. As frequency of most common bigrams and unigrams served as the second and third most important predictors in our model, we saw that the "good" responses generally focused more on specific topics that were discussed by the speaker.

We would also not recommend the use of sentiment analysis on future responses. As we discovered by reading the responses manually, as well as from our exploratory data analysis, a vast majority of these responses are bound to be neutral, since there is not much editorializing possible when reporting on what you learned from something. Since the sentiments of the responses were dominantly neutral, they did not offer much in the way of separating between responses and thus classifying responses as thorough or not thorough.

For similar data collection in the future, we would advise that the commenters are asked for more demographic information (e.g., age, gender, post-graduation plans, GPA, etc). As mentioned earlier, obtaining this information would open many avenues for further understanding and analysis of the comments left for a given speaker. Understanding the commenter themself, in context to what they wrote for each speaker, could make a model that would allow the professor to invite speakers tailored to a specific group of students.

Collecting more comments would also be beneficial. Understandably, there are only so many students in each class per quarter, but if the talks were open to other statistics students not enrolled in the course (including underclassmen), it would be advantageous to gather their comments as well. It may be interesting to compare the responses of first-years to a speaker to those of fourth-years. In a different direction, a prospective cohort study between students of different years could be interesting (i.e., comparing the career direction of first and fourth-year students that attended the same speaker after four or five years). Also potentially, analyzing the comments of non-statistics majors may unveil interesting information in terms of how a student's major may affect their thoughts on a speaker.

Sources on VADER Sentiment Analysis

- https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/
- https://github.com/cjhutto/vaderSentiment
- https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f