UCLA

Analysis of Stand-up Comedian Routines

Exploring relationships between demographics and vocabulary

Raymond Astorga
Christine Marie Castle
Jason Chhay
Byungwoo Kang
Hortencia Mendoza
Nina Nguyen

April 26, 2023

Abstract

Stand-up comedy is a comic style where a comedian performs in front of a live audience, commonly speaking directly to them. Using jokes and satire, the comedian furnishes enjoyment to an audience. The words and vocabularies chosen by comedians can come from various factors. In order to discover the relationship between the words spoken and the comedians' demographic background, we created word clouds and graphs to display any trends we found. Consequently, this revealed various insights on vocabulary usage, which relate to aspects of society such as the generation one was born or the color of one's skin.

In this report, we worked with data that consists of 369 stand up routines, which related to 158 unique comedians. We acquired these routines in the form of transcribed text, including the comedian's name, as well as the title and year of the performance. Routines typically featured introductory and concluding music, along with some comics singing in their routines. Each comic's demographic background consists of age, gender, race, country of birth, sexual orientation, net worth, birthday. Besides demographic background, additional factors were were taken into consideration. These factors include blue meter, the number of favorites, who's funnier score, wins and loses count and their biography.

We found relationships between vocabulary used and demographic background, largely based on age, gender, and race.

Contents

1 Introduction

In comedy, the goal of the comedian is to use jokes and satire to entertain an audience. Stand-up comedy requires that the comedian perform live to an audience, usually in a venue setting.

The kind of humor and words that a comedian says can come from a multitude of factors. For example, the audience that they are performing for could influence their vocabulary usage. Or, the comedian themselves has a predisposition to using certain words due to their upbringing, community, sense and style of humor, etc.

Our goal was to formally discover any relationships between demographics and vocabulary used during these kinds of routines. Demographics consist of any characteristics that a comedian has, such as their race, gender, age, etc. Vocabulary is anything that was spoken by the comedians during their stand-up routine, and this can relate to topics like swearing frequency, sentiment, reading level, and many others.

2 Data Discussion

The data we worked with consists of 369 stand-up routines. There is a total of 158 unique comedians. This data was acquired from Scraps From the Loft, a website that hosts transcripts of stand-up routines along with articles related to entertainment, such as book and movie reviews [1].

Variables that we could extract from these routines were the text from the transcript, the comedian performing, the name of the show, and the year of the show.

Since there is no demographic information about these comedians within these routines, we manually searched for these comedians online and acquired the following demographic information as potential variables to investigate: **age**, **gender**, **race**, **country of birth**, **sexual orientation**, **net worth**, **birth day, month, and year**. We also decided to include the **generation** that the comedian was born as a method of comparing groups of comedians with respect to their vocabulary.

The following data from the Dead Frog website [2], a database of comedians, was added for further demographic information: **blue meter**, **number of favorites**, **who's funnier score** along with their **wins** and **losses** count, and their **biography**. A **Flesch** reading score was added as well, which is a score that indicates the difficulty of reading a body of English text from 0-100 [3].

Finally, some of the routines had songs within them. These were indicated in the transcripts with eighth notes (♪). We created new variables to hold this text, as some are not even spoken by the comedians themselves, but rather are songs playing within the routine themselves.

3 Methodology & Findings

Our goal was to discover relationships concerning demographics and vocabulary. Word clouds provide an intuitive visualization of language without any complex statistical or mathematical background. For this reason, we opted for the use of mainly word clouds and bar charts to visually communicate our findings.

Figure 1: A wordcloud [4] of the most frequent curse [5] words used

Word	Frequency
shit	7755
fucking	7355
fuck	6530
ass	1860
dick	1669
nigga	1378
bitch	1348
motherfucker	1153
fucked	944
gay	943

Table 1: Table of top 10 "impolite" words used

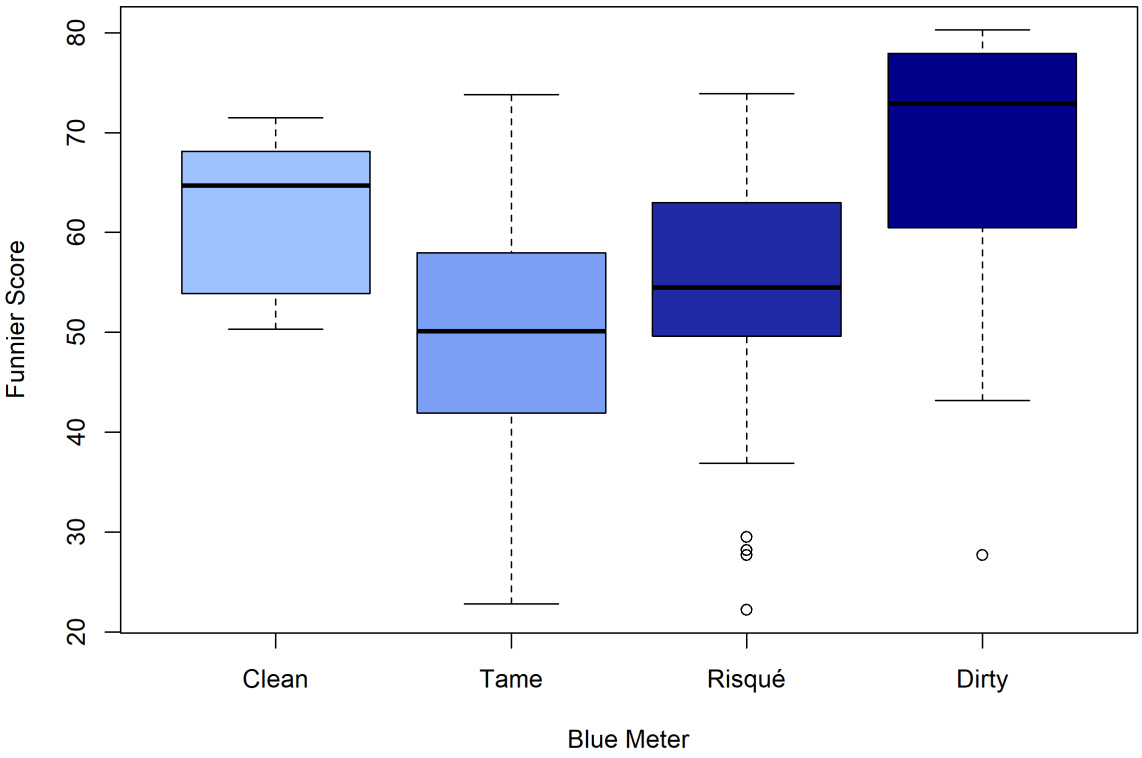

Figure 2: A boxplot of "blue meter" rating with respect to "who's funnier"

On the Dead-Frog website, each comedian has a ranking for how profane their comedy routines are: this goes from the least offensive "clean" to the most offensive "dirty". They also have a score indicating how funny they are, ranging from 0 to 100. As evident of the figure, we can see that there is a rising trend for how funny a comedian is from "tame" to "dirty". "Clean" comedians show up as being funnier than "tame" and "risqué" comedians.

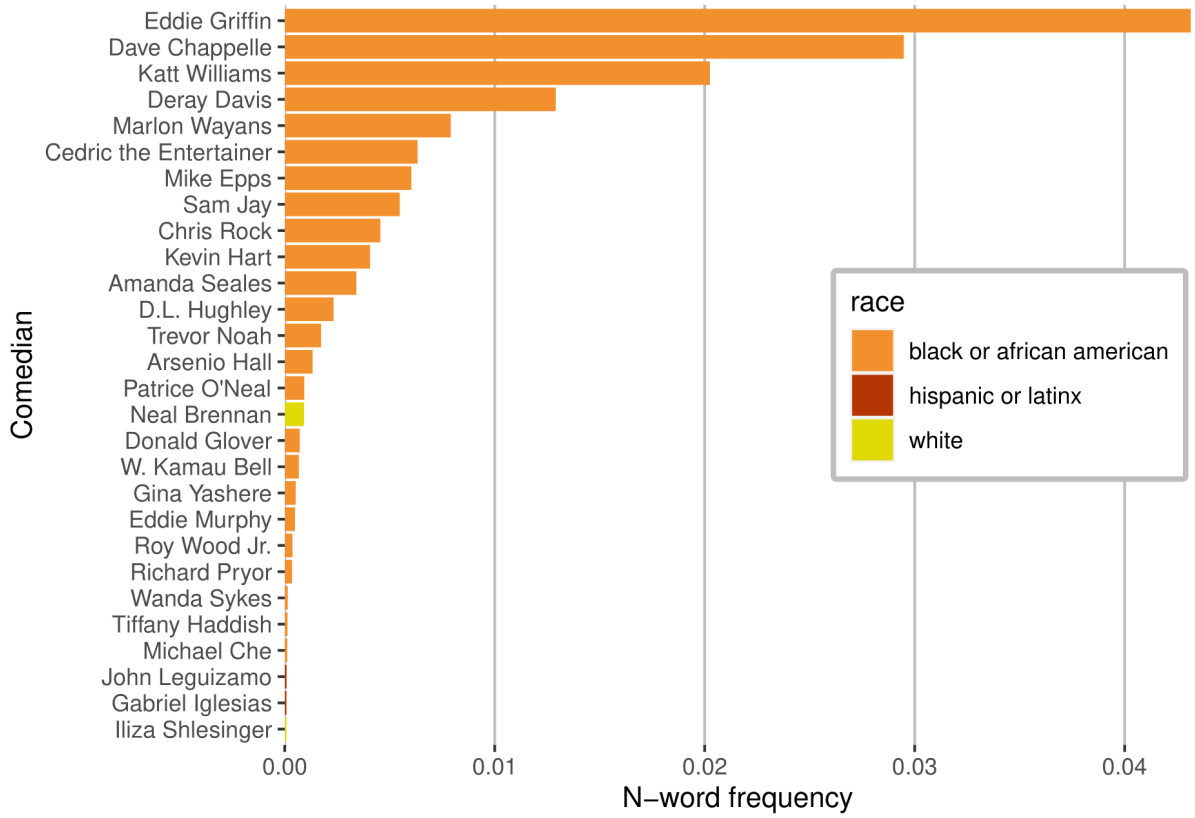

Figure 3: A chart showing which comedians say the n-word the most, with respect to race

Those who identified as black or African-American in our data were the predominant users of the n-word, with some using it more than others. Frequency is counted as the number of instances of the word divided by the total number of words spoken by the comedian. In the chart, 28 out of the total 158 comedians were found to have used the n-word in their routines.

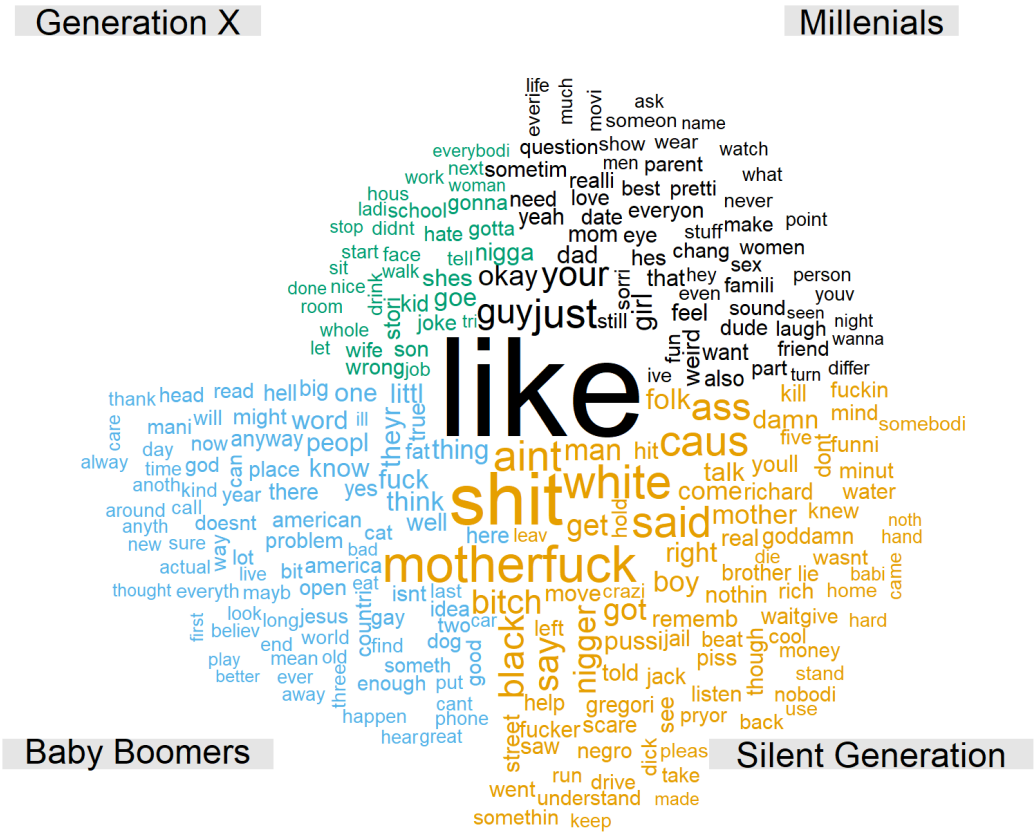

Figure 4: A wordcloud of the most frequent words, by generation

There were four generations of people found in our data. These were the "Silent Generation", "Baby Boomers", "Generation X", and "Millennials". This word cloud encompasses the frequency of the words spoken by the comedians with respect to those generations.

Observed Chi-squared value	Degrees of freedom	P-value
791207.9	2997	0

Table 2: Chi-square goodness of fit test results for the four generations

We performed a chi-square test on the words spoken most by each generation with respect to the top 1000 words spoken in all of the routines. Essentially, this test tells us if the groups are significantly different in their vocabulary. The p-value we acquired was outputted as 0, telling us that these four generations of people do significantly differ in their vocabularies.

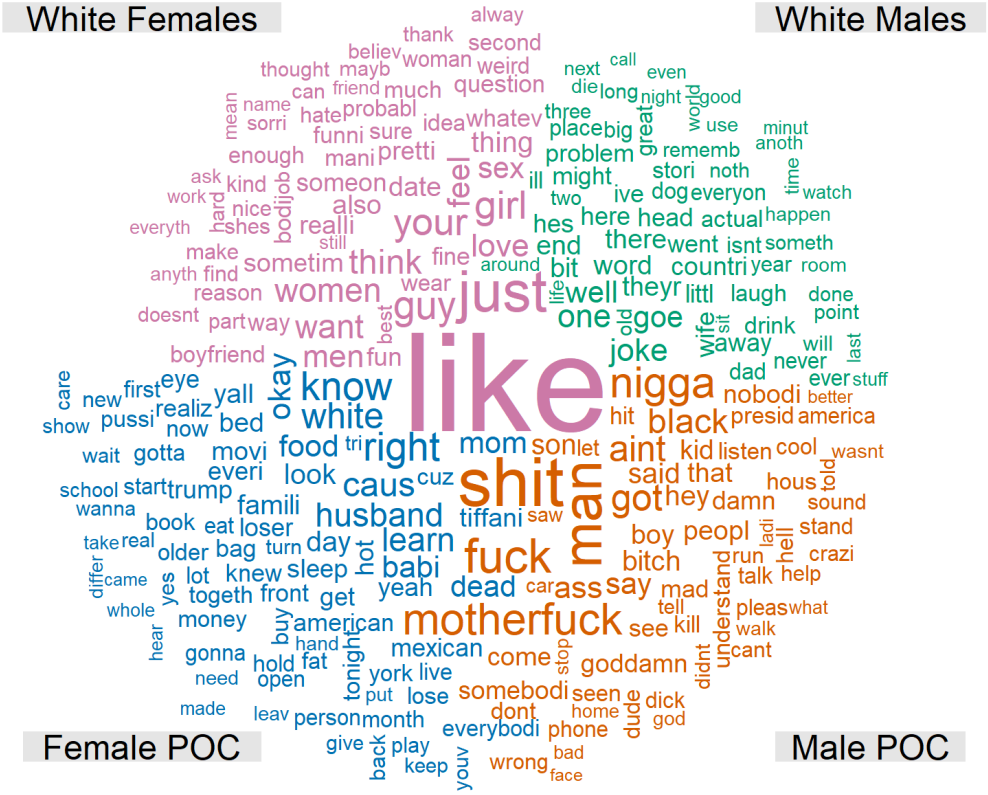

Figure 5: A wordcloud the most frequent words, by race and gender

People in our data set were categorized as female or male, and white or persons of color. This word cloud encompasses the frequency of words spoken by the comedians with respect to those categorizations.

Observed Chi-squared value	Degrees of freedom	P-value
600421.9	2997	0

Table 3: Chi-square goodness of fit test results for race and gender

Again, we performed a chi-square test to test for significantly different vocabularies. Like the test before, we acquired a p-value of 0, telling us that these four groups of comedians, separated by race and gender, also significantly differ in their vocabularies.

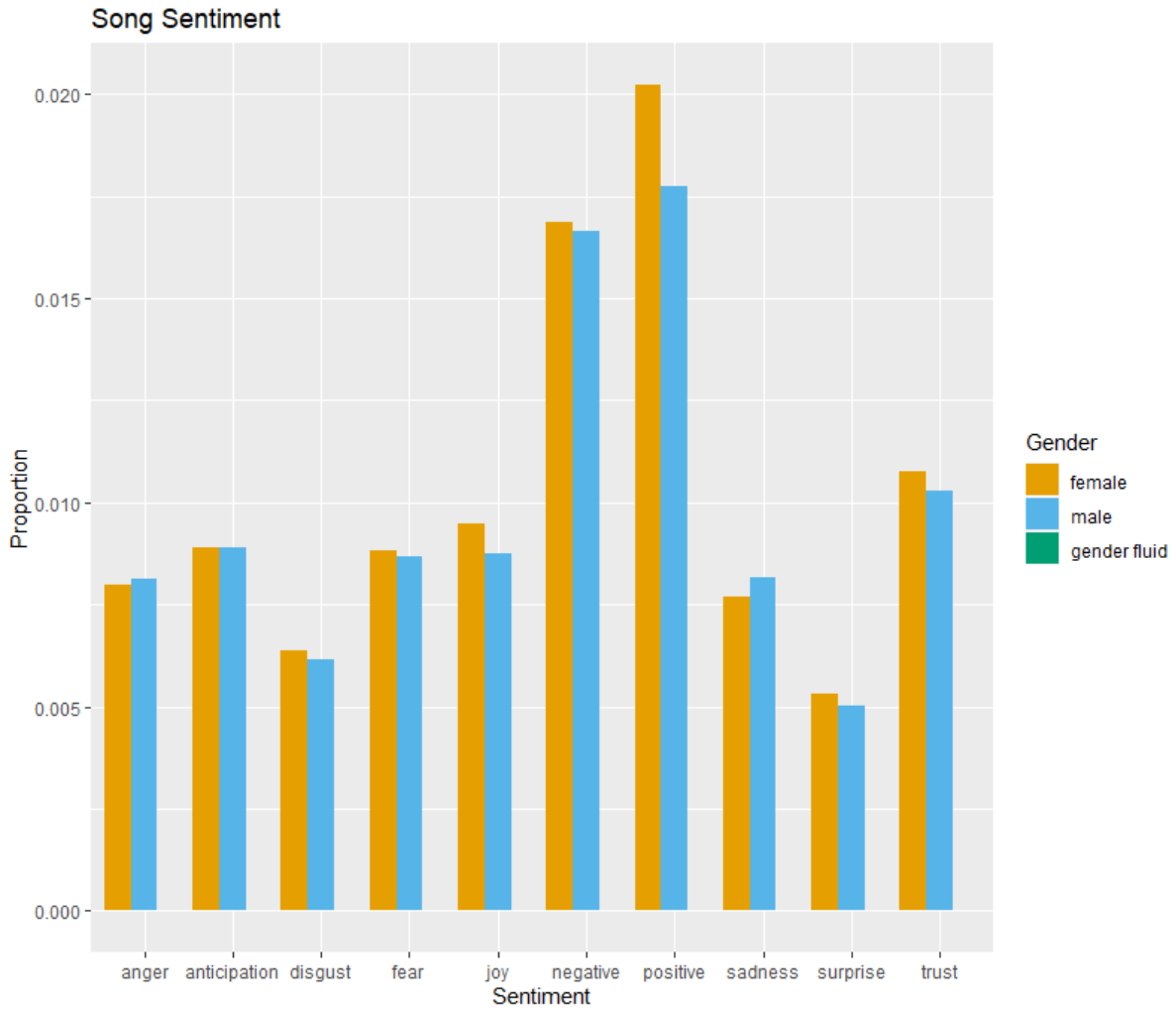

Figure 6: Emotion sentiment classification by gender

Sentiment was analyzed from introductory and concluding music, in addition to any singing a comedian performed themselves. We accounted for 3 possible gender classifications, female, male and gender fluid. Comedians who identified themselves as "gender fluid" did not have any introductory, concluding music or self performed singing included in their text transcriptions. Besides the positive sentiment, there is virtually no difference in song sentiment between those who identify as "female" and those who identify as "male".

4 Conclusion

Stand-up comedy is one of the few fields where profanity is generally accepted and widely used; in fact, we found that the "dirty" comedians are typically considered the funniest. On the other hand, "clean" comedians have the second highest average "Funnier Score". So maybe people aren't necessarily laughing at the curse words themselves, but the subversion of the norms of polite society. In the sixties and seventies, obscenity was tolerated so little that comedians would get arrested for it. Perhaps this is why words such as "shit", "motherfuck[-er/-ing]", and "bitch" are so prevalent among the comedy routines from the Silent Generation.

A stand-up's jokes are often personal and reflect who the comic is as a person, so it makes sense that a comic's identity would influence their use of profanity as well. The frequent use of male-gendered words such as "guy", "boyfriend", "husband" among female comedians could be a sign of how much the effects of a patriarchal society shape women's lives. This can also be seen in females using a higher proportion of positive song sentiments. Similarly, the frequent use of the n-word and the word "black" among male comedians of color could be a sign of how much the effects of centuries of oppression and white supremacy shape the lives of people of color, particularly Black Americans.

Stand-up comedy is an art form in which the artists use words to make people laugh. The words comics choose can reveal truths about us, from our society and culture as a whole to the individual comedians themselves.

References

[1] Scraps From the Loft, "Stand-Up Comedy Transcripts." https://scrapsfromtheloft.com/stand-up-comedy-scripts/.

[2] Dead Frog. https://www.dead-frog.com/.

[3] Wikipedia, "Flesch-Kincaid Readability Tests." https://en.wikipedia.org/wiki/Flesch%E2%80%93Kincaid_readability_tests.

[4] D. Singh, "Visualization of Text Data Using Word Cloud in R." https://www.pluralsight.com/guides/visualization-text-data-using-word-cloud-r.

[5] AllSlang, "Swear Word List & Curse Filter." https://www.noswearing.com/dictionary.