Final Project

Grant Allvin (gallvin)

December 11th, 2023

1 Introduction

Of the many ways one can learn about a community's culture, spoken conversation is something that may be overlooked in favor of things more tangible like media or objects of the time. However, by analyzing what is discussed in spoken conversation, one could learn about what events, activities, or other items and ideas that surround the community and thus define its culture. Given recent developments in conversational corpora, we are at a good time to utilize them to answer questions related to culture in this avenue.

One part of culture and community that is often of interest is the social dynamics between different demographic groups within that community. In the context of spoken conversation, discussions between those of different demographics may reveal the different ways in which the community affects them, social norms impacting their background, and perhaps even experiences shared across different groups. With this idea in mind and what is made available in the data set of interest, we aim to determine the existence of patterns in spoken conversations across different participants' sexes, races, and age groups through topic model analysis to answer the following questions:

1. How do discussion topics in spoken conversation differ based on the speakers' sexes, races, and ages?
2. Are there discussion topics we can infer to be universal, or those that may be prevalent regardless of the participants' backgrounds?

2 Data

Our data set of interest is the CANDOR corpus – a corpus of conversational American English collected from online video chats between pairs of participants and each conversation paired with its own video, audio, transcription, and some media metadata and some survey data collected after the conversation with some demographical information. This corpus includes 1656 video chats collected between January and November 2020 with participants aged 19 to 66.

For the purposes of our research questions, we split conversations on three demographics of participants: sex, race, and age group. In our data set, age is given as a number rather than an age group, so we make these age groups ourselves. With the common practice of categorizing adults based on three categories (young adult, middle-aged, senior), we split the age range into thirds. This leaves us with categories for those aged 19 to 34, 35 to 50, and 51 to 66.

For each demographic group, conversations are split between those had between participants of the same or different backgrounds. For example, when splitting on sex, we categorize conversations based on whether they are had between two males, two females, or one of each sex. With race and age group having more than two categories, it may then be useful for our questions to categorize these further into conversations had between different combinations of these groups (i.e. white and black, younger and older, etc.) Given the fewer age groups and more race groups, which may affect the representation of different combinations within these demographics in our analysis, we do this for age groups but not for race. If the conversation is missing relevant demographic information, we omit it from analyses on demographic groups they miss. In the case of race, this includes conversations with participants selecting some non-NA options, such as blank data or the "prefer not to say" option.

We note some limitations in the data set that may affect our analysis on the relevant research questions. Firstly, the data were collected on conversations had on online video chats. How we may act in online video chats may differ from how we may act in other ways, so we may not be able to generalize our findings for all methods of spoken conversation, such as phone calls or in-person conversations. Second, the data were collected in the year 2020. The year 2020 was infamously a tumultuous time in the world, and especially in the United States. With the COVID-19 pandemic, the murder of George Floyd, and an election year involving a controversial presidential candidate, events of the time may be very much unlike those in other time periods. Thus we may not be able to generalize our findings to other time periods.

We note that how we categorize our age groups may misrepresent conversations had between those close in age. For example, holding all else equal, it is likely that the experiences of a 34 year-old and 35 year-old are similar. But because they are in different age groups in our analysis, they are placed within the young-middle mixed category with other conversations with larger age gaps, which may dilute the patterns of their conversation. Also, there are considerably more conversations had between two younger people or a younger and middle-aged person than any other age category. Thus when we conduct topic analysis on other demographics, the distribution of topics may be skewed more towards those most discussed by younger people.

We note that some categories in the race demographic groups have too few observations which may misrepresent conversations had within said groups in our analysis. Namely, we have 0 conversations had between two indigenous people and 1 conversation had between two people of mixed race (not to be confused with those had between people of different races, which we label "mixrace"). Because we are either unable to analyze topics from these groups or such would not represent these groups well, we refrain from creating and comparing topic models for these groups. While there are some groups with considerably fewer observations than the white and different-race majority (black and latino with 10, asian with 36), such may be enough for our modeling purposes. But like with the groups we omit, we note that such fewer observations may misrepresent discussions had between people of the same race within these races or topics may be difficult to identify. The same applies to conversations had between two older people (17 observations).

Note that in Tables 1 to 4, token counts for demographic groups are words counted after a pre-processing procedure explained in our Methods section.

Subcorpus	males	females	mixed	total
Tokens	295,342	427,667	812,696	1,535,705
Conversations	316	451	887	1654

Table 1: Conversation and Token Count for each Sex Combination

Subcorpus	white	black	latino	asian	different races	total
Tokens	618,568	7,007	7,330	30,985	782,117	1,445,317
Conversations	664	10	10	36	838	1558

Table 2: Conversation and Token Count for each Race Combination (Omitting Multiracial and Indigenous)

Subcorpus	young	middle-aged	old	total
Tokens	456,924	155,513	16,531	628,668
Conversations	508	161	17	686

Table 3: Conversation and Token Count for Participants in the Same Age Group

Subcorpus	young & middle	young & old	middle & old	total
Tokens	540,464	196,008	123,027	859,499
Conversations	578	209	129	916

Table 4: Conversation and Token Count for Participants in Different Age Groups

3 Methods

To determine and compare discussion topics between conversations had in different demographic groups, we build structured topic models with the stm library in R for each subcorpus[3]. Within the package, this is done by creating a number of collections of words based on their frequencies, positions, and other covariates in each document that supposedly corresponds to a topic[4]. For the sake of the scope of this experiment and because finding more topics makes them harder to interpret, we set it so that we generate 5 prominent topics from the conversations.

Before we create these models, we perform some pre-processing on the text by tokenizing on singular words, decapitalizing, removing symbols, and removing words based on a given stop word list from the tidytext library and a curated one from handpicked words (See Appendix). We note that such lists of words from tidytext and ourselves are not exhaustive and may be picked based on a biased interpretation of how specific a word should be to relate to a discussion topic.

After this process, we plot the models to compare their topics, inferring each topic by looking at its words with the top 5 highest FREX values. FREX differs from the default probability metric which we used for word handpicking in that it is a harmonic mean between a word's probability rank within the topic and its rank of exclusivity to that topic. That way, it increases with a word's frequency and exclusivity in the topic. Identifying words based on frequency and exclusivity has been used in past literature to improve the interpretation of such topics[1].

4 Results

Here, we analyze topic patterns in each demographic individually before then aggregating our findings to find universal discussion topics.

4.1 Sex

When analyzing topics modeled from conversations with participants of the same or different sex, we observe some topics they share or are similar to each other. Firstly, all groups appear to share prominent topics related to domestic locations and the weather (i.e. ”portland” in males (Figure 1), ”snow” in all groups (Figures 1-3), etc.). Such topics may surround conversations where participants had asked each other where they were from, which may be a fairly common practice during a conversation. Another topic they appear to share is education, as they all have topics with ”classes” and ”semester” as prominent words. This makes sense intuitively, as education is often a shared experience among most American adults. Though because we observe that there are more conversations involving young people than any other age group, this pattern may be a result of a data skew, as it may be more likely for younger adults to use more specific education language like ”semester” while they are in or more recently left university. While there are some differences in the following, all sex groups appear to discuss entertainment and media in some form, with words like ”games” being found in male and mixed conversation topics and ”movies” being found in a topic for all groups. This may imply that entertainment is possibly a universal item enjoyed by most Americans and thus a notable part of American culture.

We also observe some other similarities which may be affected by the time period the data were collected. Conversations between two males and those of different sexes (Figures 1 and 3) share topics related to surveys and other side hustles, both having topics with ”survey” and ”mturk” (an outsourcing service by Amazon) as most prominent words. Such a topic being absent in female conversations may be due to either the topic being less prominently discussed among females or is motivated by the presence of males. Either way, its presence in our models may be attributed to the consequences of the COVID-19 pandemic in the United States in 2020. Due to the economic shifts motivated by the practice of staying home, people had lost their jobs or otherwise had financial trouble. Perhaps to make some more money while stuck at home, they took up taking surveys and doing work on MTurk. This may be a practice some people do outside of this time period, but given the financial circumstances of that year, it may have been more popular then and perhaps lost popularity after the pandemic.

Conversations between two females and those of different sexes (Figures 2 and 3) share topics related to politics, both having topics with ”voting” as a prominent word among other words related to politics (”trump”, ”biden”, etc. in females and ”americans” in different sexes). Similar to surveys, this topic does not appear to be prominent among conversations between two males. Though one of their topics includes ”trump”, but it’s grouped with other words related to movies, so it might just not be prevalently discussed topic for males or is motivated by the presence of females. This topic’s prevalence may be motivated by 2020 being an election year, and one with a rather controversial presidential candidate at that with former President Donald Trump running again for a second term. While it may be the case that politics may be discussed more often at every election year, attention to politics may be further inflated by 2020’s political circumstances.

While all groups appear to share some similar discussion topics, there appear to be some differences which may highlight social norms dictating standards of the sexes. Firstly, while ”games” is a word shared by topics between two males and those between different sexes, ”games” in male conversations are grouped with words related to video games, such as ”xbox”, ”pc”, and ”rpg”. This being a topic more prevalently discussed among males may be due to video games in the mainstream being considered a more ”masculine” hobby or social stigma against women playing video games. For topics among two female participants (Figure 2), the shared ”movie” word is also grouped with ”netflix” and ”episode”. This being related to TV and being absent in other sex combinations may then imply that television is a hobby enjoyed more among females.

A topic fairly remote from the differences above that appears to be exclusive to females are pets, with one topic including prevalent words like ”cat” or ”dogs”. Such may be discussed more among females due to them caring more about and for their pets than males do, which may then be motivated by social norms related to evolution that lead to females being more nurturing[2].

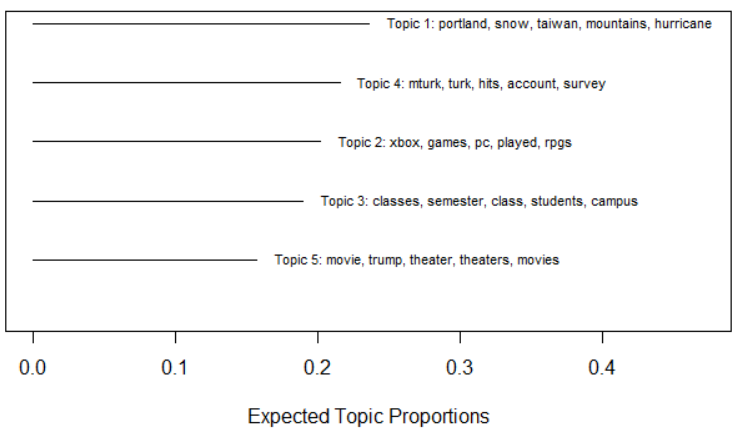

Figure 1: Discussion Topics Between Two Males

4.2 Race

For identifying patterns involving race, we find it difficult to use the lines of logic we had used for sex to determine whether people of different race combinations discuss similar topics, given the small number of conversations had between two people of the same minority race. While we may say that a topic being present in one sex and mixed sexes but not in all sexes may be due to the topic being motivated by the

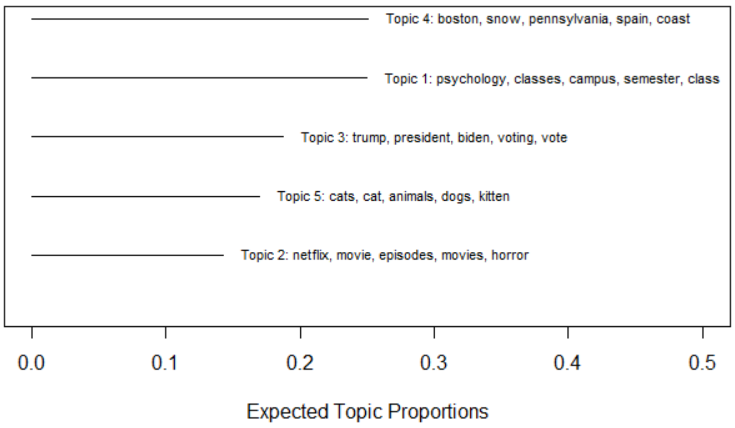

Figure 2: Discussion Topics Between Two Females

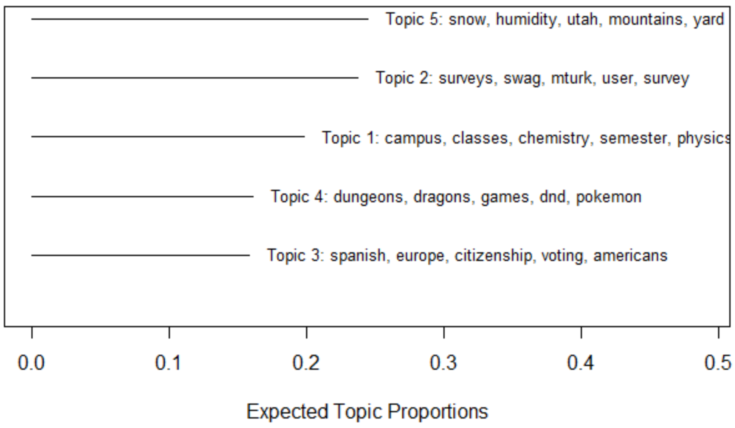

Figure 3: Discussion Topics Between People of Different Sexes

presence of some sex, we find it unsound to analogize this to saying, say, a topic discussed by two white people and those of different races but not by two minorities of the same race is influenced by the presence of white people. So we instead attribute topics under this condition (which are rather common) to being shared among all races and keep in mind the limitations of doing so.

Similar to when we analyze on gender, those of most race groups appear to share topics related to education (i.e. ”semester” in topics between two white people (Figure 4) and those of different races (Figure 8)), surveys (which we may extend to finances, as while white and different race conversations share similar words like ”mturk”, ”surveys”, and ”swag” (likely referring to the survey webiste Swag Bucks), conversations among two black people in our data appear to have topics with the words ”bank” and ”marketing” (Figure 5)), and locations and the weather (i.e. ”francisco” in white conversations, ”houston” in black conversations”, and ”wyoming” in hispanic/latino conversations (Figure 6). While words related to locations are also found in discussion topics between two Asian people (Figure 7), they do not appear to be grouped with enough words to compose a notable topic, which may be due to how few conversations there are in this race category. Nevertheless we make the assumption that location is a shared topic among races.

We identify pets as a prevalent topic among all races as well, with conversations between white people and different races having topics with the shared words ”dog” and ”cat” among others.

Interestingly, a prevalent topic we identify in conversations between people of different races, but not between those of the same race, is entertainment, with words like ”brady” (maybe referring to Tom Brady), ”mario” (likely referring to Super Mario), and ”xbox” being highly prominent. This topic also contains the word ”game”, which is also found in a conversation topic between two white people, but with it being grouped with more survey words in the latter category leads us to believe entertainment is not nearly as prevalent of a topic for the latter category. This pattern may occur due to entertainment being, once again, a universal item of American culture.

4.3 Age

Like with race, we find it difficult to make observations of patterns on some conversation categories split on age group combinations, given a low number of conversations in said categories. Namely, this issue mainly occurs with conversations between two older adults, which we recall has 17 conversations. Note that because of this, we do not include a topic model figure for older adult conversations.

Similar to topic models composed on previous demographic groups, we observe that all age groups appear to often discuss politics (i.e. ”trump” in conversations between two young people, two middle-aged people,

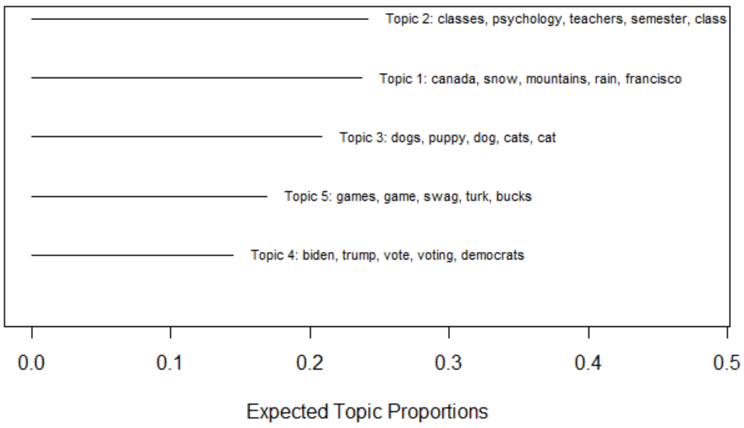

Figure 4: Discussion Topics Between Two White People

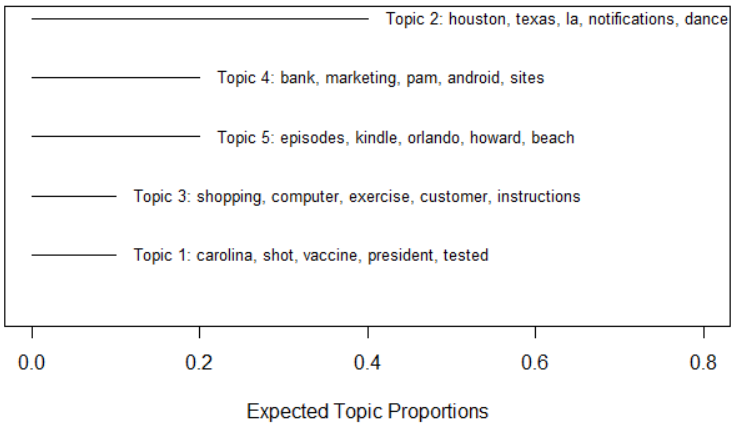

Figure 5: Discussion Topics Between Two Black People

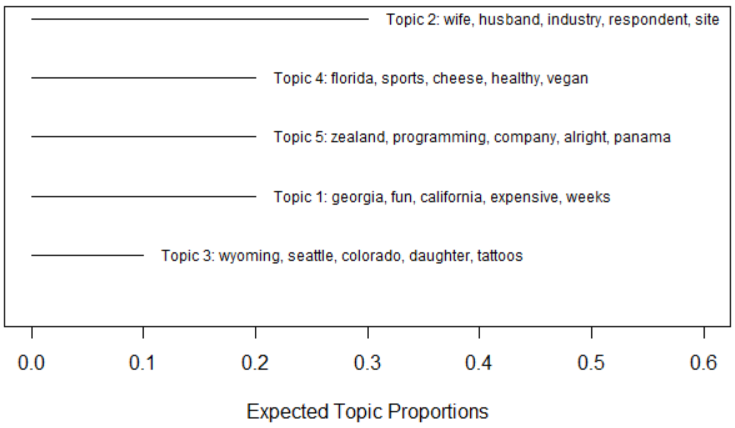

Figure 6: Discussion Topics Between Two Hispanic/Latino People

and a young person and an old person (Figures 9, 10, and 12 respectively)) and entertainment (”games” found between two younger people and a young person and old person, ”movies” found between two middle-aged person or a middle-aged person and a young person (Figure 11), and ”music” found between a middle-aged person and an old person (Figure 13)). Like with race, pets are also a prominent topic among all age groups (”cats” among other words found between two young people, a young and middle-aged person, and a middle-aged and old person).

Interestingly, while other demographic splits developed models with topics related to location, much of them only extended to North America and may only serve the purpose of learning where someone else is from. On the other hand, location topics built on age groups appear to be more international, with words like ”europe” in young people’s conversations, ”ireland” in middle-aged people’s conversations, and ”german” between young and old people. Thus a topic related to travel may be something that is identified here.

While all age groups appear to discuss school to some extent, the language used to talk about school may

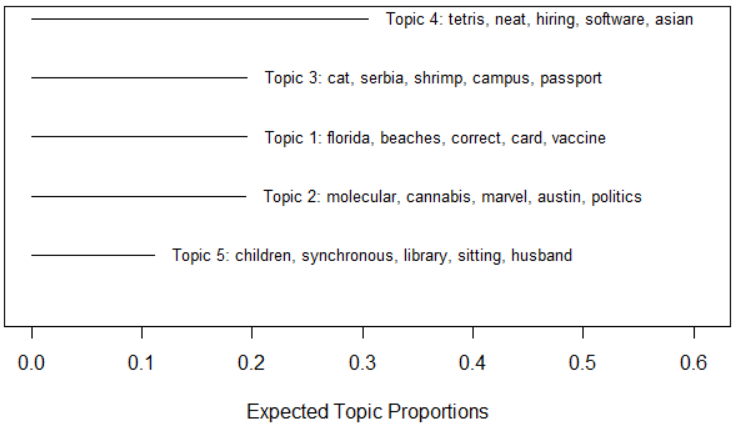

Figure 7: Discussion Topics Between Two Asian People

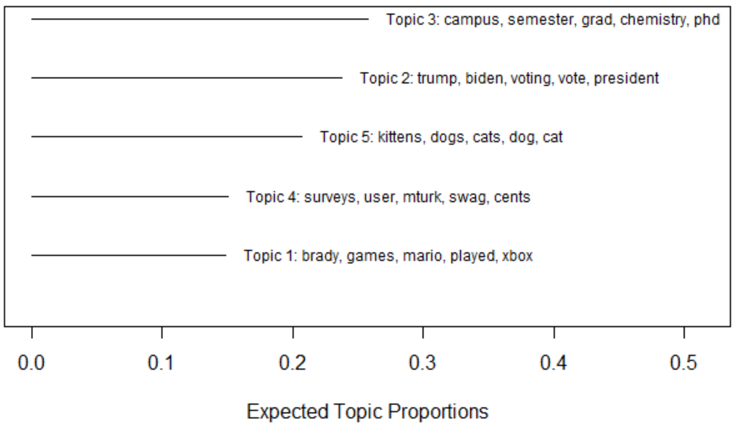

Figure 8: Discussion Topics Between People of Different Races

differ by age. In particular, in conversations that involve young people, the topics related to education include words that imply active involvement in education, such as disciplines (”physics”, ”chemistry” between two young people) and academic locations (”campus” between a young and middle-aged person, ”classroom” between young and old people). This makes intuitive sense, as this age group is comprised of people who are either still in education or have left school more recently. Conversations involving middle-aged and older people may have these patterns as well, but they are often accompanied by the word ”kids”, implying they are discussing more about their children in school rather than themselves. Such differences within the same topic may then illustrate education as a virtually universal experience among adults, but additionally highlight how one’s involvement in education (which is often related to their age) can affect the ways in which education is discussed.

Among age groups, it appears that the sub topic of video games in the area of entertainment is one that is exclusively more prominent in discussions between young people, as they have a topic describing gaming systems (i.e. ”xbox”, ”ps” (PlayStation), ”pc”). This being more prevalent in this age group and absent in all other combinations may imply that video games are often a younger person’s hobby, or that this is a consequence of video games being more popular and accessible while or before they grew up.

Interestingly, a topic that appears to show up in conversations involving middle-aged people is surveys and other side hustles, as observed by topics including words like ”testing” and ”user” between two middle-aged people, and ”mturk”/”turk” and ”survey” found between middle-aged people and the other two age groups. Such a topic being more prevalent when middle-aged people are involved may be a consequence of this age group often being the most employed, at least before the pandemic. Much like why this topic is prevalent on the basis of sex, the economic shifts brought about by the COVID-19 pandemic in the United States led to higher rates of unemployment or otherwise financial troubles. To make some extra money while forced to stay home, those in the work force - which most middle-aged people likely are - may have turned to taking surveys and doing tasks on MTurk during the pandemic. It may be the case then that as the pandemic ended, this activity has decreased. Thus while surveys and other side hustles may be a prominent topic discussed among middle-aged people in our data, such may not be generalizable to other time periods.

4.4 Universal Topics

With the most prominent topics discussed between people of the same or different demographic backgrounds identified, we then identify which ones were found among all sexes, races, and age groups and their combinations to then identify universal topics that may define the relevant culture. While there are some topics

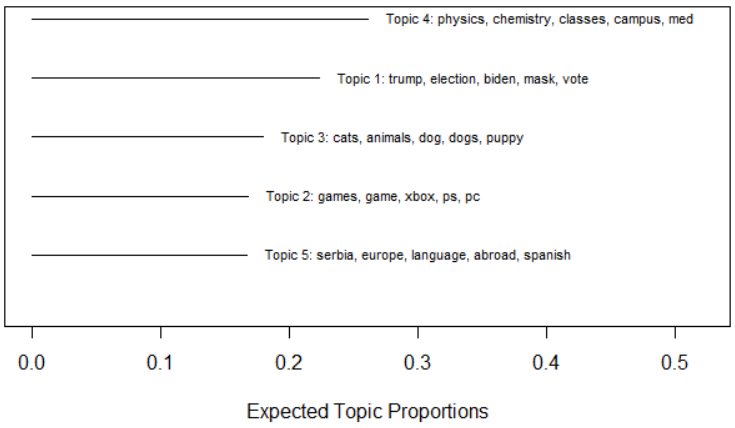

Figure 9: Discussion Topics Between Two Young People

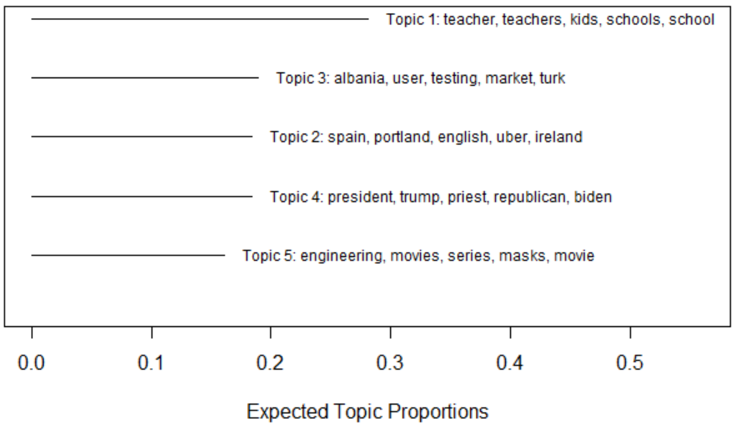

Figure 10: Discussion Topics Between Two Middle-Aged People

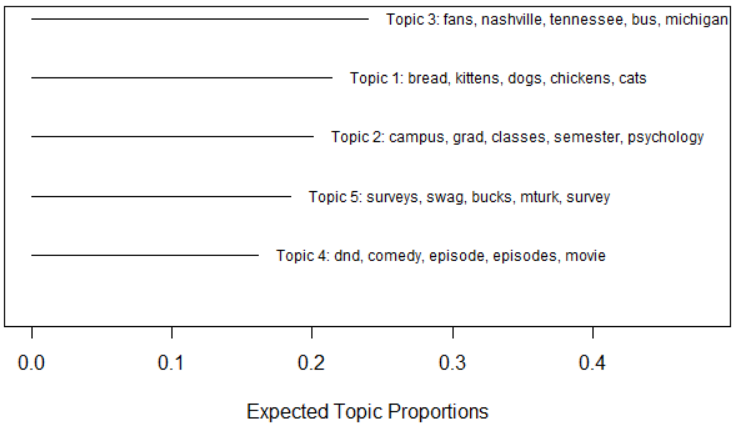

Figure 11: Discussion Topics Between a Young and Middle-Aged Person

found to be shared among one or two demographic groups such may not be considered universal because some subcategory within some other group did not have said topic as a prevalent one. With that in mind, from our model comparisons and the assumptions we make when faced with limited data, we consider **education, entertainment, and places** to be discussions topics that are universal and thus may be considerably definitive of the culture the participants are in, at least for the year the data were collected.

5 Discussion

From our findings, we infer that discussion topics in the given data considered universal, as defined by being prevalent in all (or most from our assumptions of race and age due to small data) sexes, races, and age groups, are education, entertainment, and places. In our analysis of each demographic group, we observe some differences among groups that may either be consequences of social norms or the given time period.

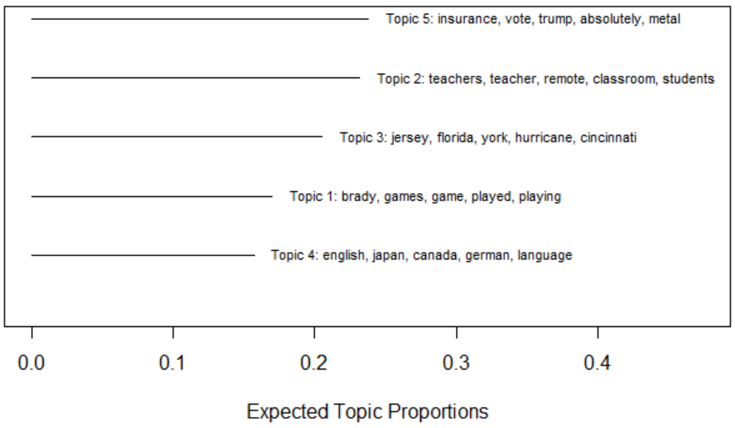

Figure 12: Discussion Topics Between a Young and Old Person

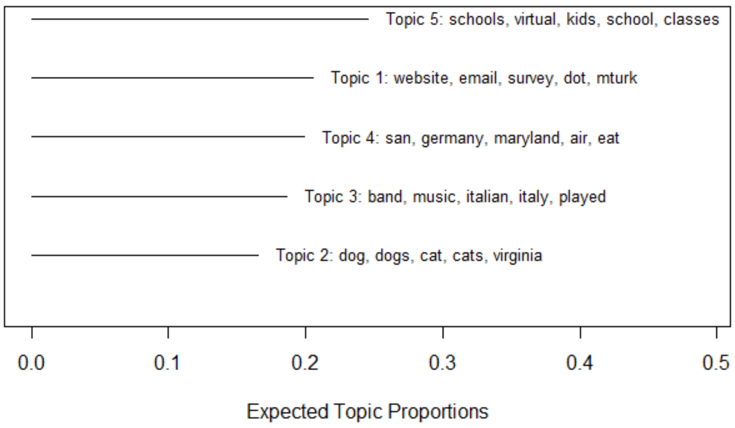

Figure 13: Discussion Topics Between a Middle-Aged and Old Person

Notably, video games were discussed more among men and younger people. The former has implications of video games being considered a more ”masculine” hobby in combination of social stigma against women playing video games. The latter has implications that video games are also a hobby most preferred by young people, which may be a consequence of its growing popularity and accessibility coinciding with or coming before them growing up. Topics related to television and pets being exclusively prevalent to women may also be consequences of television being considered a hobby enjoyed more by women and the social norms for women to be nurturing motivated by evolution respectively. Due to smaller subsets of data for conversations between same-race minorities, we are unable to determine whether topics may change based on race. On age, we determined that surveys and other side hustles may be more prevalently discussed among middle-aged people, but this may be a consequence of financial troubles brought about by the COVID-19 pandemic during which the data were collected.

Whether these observations can be truly generalized to all of spoken English in the United States is debatable. As stated before, these transcriptions were collected in the year 2020, a time period very much unlike any other year. Particularly when we identify topics related to politics and side hustles, it may be the case that the prevalence of topics are more biased towards that year’s events and thus may not apply to today. Additionally, it may be the case that people act differently on video chats versus face-to-face or phone conversation. Something else of note is that because participants were randomly sampled to participate, it is likely that conversation partners did not know each other before. It may be the case that discussion topics between friends or those in other close relationships may differ, so our patterns may not be generalizable to all dialogue in different relationships.

We fell into a few pitfalls due to the constraints of our data and our methods which limit our ability to make observations. Like what we had done with sex, we were hoping to have enough data to identify cohesive topics that may be exclusive in conversations between same-race minorities. But we had considerably few, which made it difficult to identify some topics. The same applies for older people. Perhaps in future study, conversations among more same-race minorities could be collected for topic analysis. Additionally, how we categorized participants by age group may hide some patterns between people who are close in age but considered to be in different age groups. A potential solution to work around this may be to make more, smaller age groups, but we decided against this for the sake of our limited scope in this project.

References

[1] J.M. Bischof and Edoardo Airoldi. “Summarizing topical content with word frequency and exclusivity”. In: *Proceedings of the 29th International Conference on Machine Learning, ICML 2012* 1 (Jan. 2012), pp. 201–208.

[2] David G. Rand et al. “Social heuristics and social roles: Intuition favors altruism for women but not for men.” In: *Journal of Experimental Psychology: General* 145.4 (Apr. 2016), pp. 389–396. ISSN: 0096-3445. DOI: 10.1037/xge0000154. URL: http://dx.doi.org/10.1037/xge0000154.

[3] Margaret E Roberts, Brandon M Stewart, and Dustin Tingley. “Stm: An R package for structural topic models”. en. In: *J. Stat. Softw.* 91.2 (2019).

[4] Margaret E. Roberts et al. “The structural topic model and applied social science”. In: *International Conference on Neural Information Processing*. 2013. URL: https://api.semanticscholar.org/CorpusID:59893873.

A Appendix

Besides the list of stop words provided by the tidytext R library, the following words were handpicked from the corpus data to be removed when developing our topic models:

”yeah”, ”uh”, ”mhm”, ”um”, ”huh”, ”people”, ”time”, ”lot”, ”stuff”, ”pretty”, ”gonna”, ”mm”, ”nice”, ”day”, ”wow”, ”cool”, ”prolific”, ”guess”, ”feel”, ”bad”, ”couple”, ”minutes”, ”basically”, ”bit”, ”talking”, ”love”, ”10”, ”hmm”, ”crazy”, ”shit”, ”started”, ”week”, ”person”, ”life”, ”sort”, ”weird”,”guy”, ”talk”, ”true”,”telling”,”guys”,”steps”,”supposed”,”start”,”reason”,”understand”,”hour”,”yep”,”heard”, ”18”, ”luckily”,”hey”, ”exact”, ”ideal”, ”wait”, ”restart”, ”hear”, ”heard”, ”days”, ”gray”, ”interact”, ”task”, ”scam”, ”called”

Figure 14: Stop and Filler Words Handpicked for Removal from our Corpus Data