Unsupervised Text Analysis in Applied Settings: A Case Study in Selective College Admissions

A.J. Alvero

Unsupervised Text Analysis in Applied Settings: A Case Study in Selective College Admissions

AJ Alvero, Student Narrative Lab, Stanford University ajalvero@stanford.edu

Overview

Machine learning (ML) and AI are now commonly used in educational research for learning analytics, intelligent tutoring, and automated feedback. However, this work does not always account for the different social contexts and experiences that affect learning and schooling for students, such as race, class, and language. Understanding and analyzing these sociocultural contexts of linguistic/textual production is important to prevent the reproduction of colorblind racism in code. **ML and AI can do more for students than assess their performance by instead observing for pattern and bias.**

To highlight this issue, we compared results from vanilla latent Dirichlet allocation (LDA) and structural topic modeling (STM) on a corpus of college admissions essays (CAE) written by Latinx students. Vanilla LDA and STM generate topics with the same or similar top terms, but STM shows how the same topic varies with income, gender, and other covariates. This is evidence that the CAE is biased and should be revaluated.

Procedures

Essay Prompts and Procedure

First year applicants had to respond to two prompts:

Prompt 10	Prompt 12
"Describe the world you come from — for example, your family, community or school — and tell us how your world has shaped your dreams and aspirations."	*"Tell us about a personal quality, talent, accomplishment, contribution or experience that is important to you. What about this quality or accomplishment makes you proud, and how does it relate to the person you are?"*

Preprocessing

All stopwords and punctuation were removed; all tokens were lowercased and stemmed using the Porter algorithm. The ldatuning package in R was used to find an optimal number of topics.

Data

The data are every Latinx CAE submitted to selective universities in 2015 and different metadata.

Study Corpus: Applications Submitted Fall 2015 (To Begin College in 2016)	
Description	All self-identified Latinx applicants
Number of applications	39,667
Total documents	79,317
Prompt 10	39,663
Prompt 12	39,654
Average Word Count: Prompt 10	490
Average Word Count: Prompt 12	424

Descriptive Statistics	
Average Reported Household Income	61,752
Average Number of People in Household	4.48
Female	24,133
Male	15,303
Decline to State/No response (Gender)	231
DE Latino	39,598
Cuban	507
Latino	9,614
Mexican	30,796
Puerto Rican	746
Other Latino	4,156

Applications by Campus	
Campus 1	13,372
Campus 2	15,053
Campus 3	21,445
Campus 4	19,868
Campus 5	10,316
Campus 6	16,907
Campus 7	16,674
Campus 8	18,794
Campus 9	15,158

Results/Discussion

Vanilla LDA found semantically coherent topics, but STM shows that the prevalence of different topics are correlated with social and applicant characteristics, such as household income and campuses applied to. This points to potential bias in the CAE. This descriptive work is the first step towards more systematic analysis of CAE.

Vanilla LDA

Finding 1

Topic Number	Dirichlet Parameter	Top Terms per Topic
3	0.02836	people mexican hispanic white stereotypes race skin color american society culture black ethnicity minority latino background latina stereotype racial prove
32	0.04912	united states mexico country years u.s family life undocumented born parents opportunities back immigrant dream american border living english opportunity

Finding 1: Latinx students write about social issues specific to Latinx communities.

Finding 2

138	0.09105	family parents life school make hard college good give support work successful future things brothers proud made sisters siblings person
76	0.07912	mother family father life support mother's single home provide siblings age children growing child education young household parent raised due

Finding 2: Latinx students write about their families in CAE.

Finding 3

73	0.0586	community service helping volunteer volunteering people back local time helped give giving involved work volunteered organization hours part center experience
74	0.01717	social rights voice people issues change justice women world fight gender society human equality injustice movement speak stand awareness feminist

Finding 3: Latinx students mention social and political activism in their CAE.

STM

All Schools	Covariate Group	Prompt	Statistically Significant Covariates
Topic 53: Highest Prob: hispan, background, latino, stereotyp, minor, prove, ethnic FREX: hispan, latino,stereotyp, minor, predomin, background, label Lift: countrymen, education-less, ingenium, latino, responsibilities.a, hispan, non-latino Score: hispan, stereotyp, latino, background, ethnic, minor, race	Ethnic/racial	10	Latino: Positive correlation Mexican: Positive Correlation
Topic 89 Highest Prob: us, worri, bill, afford, stress, pay, rent FREX: bill, rent, minimum, paycheck, nlec, worri, nephew Lift: workhors, hard-hit, payday, struggle-fre, could.in, payment, evict Score: bill, rent, us, afford, pay, worri, wage	Ethnic/racial	10	Cuban: Negative correlation Other: Negative correlation Puerto Rican: Negative correlation Mexican: Positive correlation
Topic 30 Highest Prob: struggl, challeng, difficult, face, obstacl, situat, overcom FREX: obstacl, difficult, overcom, face, difficulti, challeng, struggl Lift: persus, overcom, obstacl, difficulti, reality.mi, setback, difficult Score: obstacl, overcom, challeng, struggl, difficult, face, situat	Ethnic/racial	10	Cuban: Negative correlation Other: Negative correlation Puerto Rican: Negative correlation Mexican: Positive correlation

Finding 1: Mexican students were more likely to write about social issues.

All Schools	Covariate Group	Prompt	Statistically Significant Covariates
Topic 31 Highest Prob: dad, see, start, got, rememb, went, go FREX: dad, stepmom, wallet, caller, father-daught, i, phone Lift: 3.97, dogg, older-broth, other.i, dad, 24-pack, family-h Score: dad, got, stepmom, start, phone, rememb, see	Home	10	Income: Negative correlation
Topic 46 Highest Prob: sister, sibl, younger, care, babi, twin, watch FREX: sister, sibl, twin, eldest, younger, niec, babi Lift: mervyn, moniqu, sister, arac, priscila, stork, twin Score: sister, sibl, younger, babi, twin, care, eldest	Home	10	Income: Negative correlation Gender (male): Negative correlation Family size: Positive correlation
Topic 23 Highest Prob: parent, brother, sister, older, younger, sibl, care FREX: brother, sibl, sister, oldest, parent, older, younger Lift: arlen, val, oldest, s.o.a.r, sibl, brother, eldest Score: parent, brother, sister, sibl, older, younger, care	Home	12	Income: Negative correlation Gender (male): Negative correlation Family size: Positive correlation

Finding 2: Lower income students are more likely to write about their family.

All Schools	Covariate Group	Prompt	Statistically Significant Covariates
Topic 3: Highest Prob: communiti, issu, awar, societi, youth, voic, advoc FREX: gang, advoc, injustic, discrimin, gsa, undocu, racism Lift: chirla, aclu, ageism, artiv, ayotzinapa, caravana, carecen Score: communiti, latino, youth, immigr, issu, gang, advoc	Campus	12	Campus 1: Positive Campus 9: Positive Campus 2: Positive Campus 8: Positive
Topic 89 Highest Prob: support, role, exampl, model, guid, encourag, advic FREX: support, model, exampl, role, guidanc, guid, advic Lift: professionalist, unbenefici, 17-years-old, pfs, model, support, guidanc Score: support, role, model, exampl, guidanc, guid, advic	Campus	10	Campus 1: Negative correlation Campus 9: Negative correlation Campus 5: Positive correlation Campus 7: Negative correlation Campus 6: Positive correlation
Topic 13 Highest Prob: accept, true, word, hide, embrac, reject, ident FREX: gay, judgment, hide, judgement, shame, secret, reject Lift: caitlyn, lgbt-straight, sagittarius, transphobia, homosexu, genderfluid, siento Score: accept, gay, sexual, reject, hide, secret, true	Campus	12	Campus 1: Positive correlation Campus 9: Positive correlation Campus 5: Negative correlation Campus 6: Negative correlation

Finding 3: Students applying to campuses Known for activism are more likely to write about social/political activism.

References available upon request