Exploratory Text Analysis of the “Sexual Harassment in Academia” Dataset

Zoë Wilkinson Saldaña (zoews) | April 16, 2018

1. Motivation

I learned about the “Sexual Harassment in Academia: Results of a Crowdsourced Survey” in SI 649, where the data was presented as a an option for our semester-long group project. At our first group meeting, we pulled up the dataset and discovered a messy, unwieldy collection of data about an extremely important topic. Our group surveyed the 3,000+ lines of survey responses and decided that it would be too time-intensive to extract clean data that we could subsequently visualize. We decided on another dataset.

However, the critical data scientist in me was hooked. As somebody who is passionate about using data to investigate social issues, I appreciated the importance of this unprecedented data. Thousands of individuals had revealed their experiences with sexual harassment and assault in academia, many apparently for the first time. These reports detailed the devastating effect these events had on their lives. This is vital, powerful data that deserves to have its story told.

At the same time, the issues with preprocessing and the unpredictability of free text responses contribute to the idea that we should wait for *experts* or *well-funded research teams* to dig into this data. I am eager to explore ways of doing data science that circumvent barriers by iterating quickly to generate important questions. I broke my project down into **four main questions:**

1) What are the main themes (or topics) in reports of harassment/assault event?

2) How do the descriptions of the aftermath of incidents differ between professor/faculty/department chair perpetrators and non-faculty perpetrators?

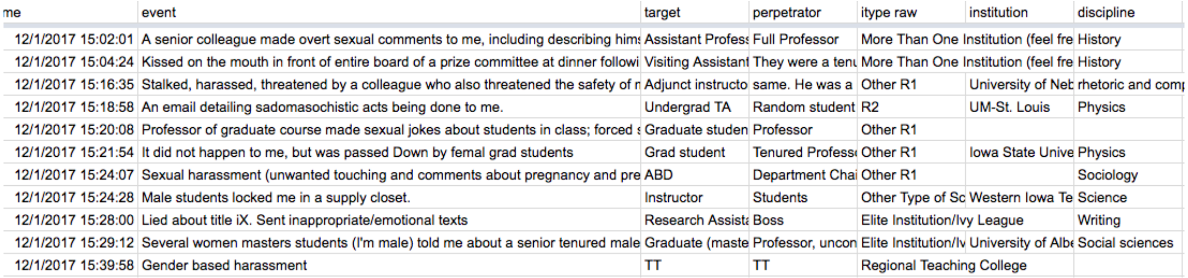
3) What language in incident descriptions is most predictive of whether the perpetrator was a professor/faculty/department chair or not? How powerful is this prediction?

4) Finally, what language in the incident report is most predictive of whether the respondent “left” or “quit” their position/academia/the discipline/etc? How powerful is this prediction (in general and relative to the professor model)?

2. Data Source

The Sexual Harassment in Academia dataset is a collection of responses to an open Google Survey. For my analysis, I download the data on April 4, 2018, when **2,438 reports** had been submitted **between December 1st, 2017 and April 4, 2018:**

- Here is a direct link to the raw data (as a Google Sheet): https://docs.google.com/spreadsheets/d/1PfaDunV1ZS-AK0neJ2SyAM4U3fajToMbP0ZYW9O6Ueg/edit?usp=sharing
- And here is a website providing a web interface for accessing the data, also linking to the original survey, a blog post by its creator, etc.: http://people.csail.mit.edu/karger/Exhibit/Harass/



me	event	target	perpetrator	itype raw	institution	discipline
12/1/2017 15:02:01	A senior colleague made overt sexual comments to me, including describing hims	Assistant Profess	Full Professor	More Than One Institution (feel fre		History
12/1/2017 15:04:24	Kissed on the mouth in front of entire board of a prize committee at dinner followi	Visiting Assistant	They were a tenu	More Than One Institution (feel fre		History
12/1/2017 15:16:35	Stalked, harassed, threatened by a colleague who also threatened the safety of r	Adjunct instructo	same. He was a	Other R1	University of Neb	rhetoric and comp
12/1/2017 15:18:58	An email detailing sadomasochistic acts being done to me.	Undergrad TA	Random student	R2	UM-St. Louis	Physics
12/1/2017 15:20:08	Professor of graduate course made sexual jokes about students in class; forced s	Graduate studen	Professor	Other R1		
12/1/2017 15:21:54	It did not happen to me, but was passed Down by femal grad students	Grad student	Tenured Profess	Other R1	Iowa State Unive	Physics
12/1/2017 15:24:07	Sexual harassment (unwanted touching and comments about pregnancy and pre	ABD	Department Chai	Other R1		Sociology
12/1/2017 15:24:28	Male students locked me in a supply closet.	Instructor	Students	Other Type of Sc	Western Iowa Te	Science
12/1/2017 15:28:00	Lied about title iX. Sent inappropriate/emotional texts	Research Assista	Boss	Elite Institution/Ivy League		Writing
12/1/2017 15:29:12	Several women masters students (I'm male) told me about a senior tenured male	Graduate (maste	Professor, uncon	Elite Institution/Iv	University of Albe	Social sciences
12/1/2017 15:39:58	Gender based harassment	TT	TT	Regional Teaching College		

The Google Sheets shown above) includes both the raw survey results and a few additional columns that attempt to clean or otherwise aggregate. However, I chose to focus only on the original raw data which reflects the actual submissions via the Google Survey, as I wanted to have the most flexibility for my NLP work.

The survey includes only two categorical variables, presented as multiple choice options: **gender of perpetrator** (woman/man/non-binary/unsure/various/other__) and **type of institution** (small liberal arts college/Elite Institution/Ivy League/Other R1/etc.) Otherwise, responses are all open text fields, and many are optional (or accept a blank or "none" as submission). These **text field submissions** include:

- Event description, target status (e.g. student, professor), perpetrator status (e.g. professor/chair/fellow graduate student/etc.), name of the institution (optional), field discipline, institutional response (if any), institutional/career consequences for perp (if any), impact on your career, impact on your mental health, impact on life trajectory/choices, and other.

I decided to narrow my focus a little in order to focus on four main aspects of these reports: **status/role of perpetrator**, **event description**, **outcome of the event for perpetrator** (response and punishment), and **impact on target** (mental health, career, and life impacts).

3. Methods

Before starting to answer my questions, I had to clean and standardize the textual data. I made all text lowercase, got rid of unnecessary punctuation, etc. I quickly encountered an issue with nulls - first off, I had to explicitly convert empty or blank spaces to Null values to be recognized as such in the DataFrame. However, much of the data had at least a field or two missing, so I could not simply drop all rows with any nulls. I decided to drop rows with some nulls as needed going forward - within each function, I generate a new dataframe and drop rows with nulls on the columns I wish to focus on.

I used the Databricks platform for all of my code. My laptop had issues with running Spark consistently, and I wanted to leverage the quicker processing time of the Databricks cluster.

1. **What are the main themes (or topics) in reports of harassment/assault event?**

I decided to approach the question of identifying themes as a topic modeling question, which I tackled with Latent Dirichlet allocation (LDA). I first constructed a PySpark MLLib pipeline that consisted of (1) tokenizing all words within a particular column, such as "event" description, (2) removing stopwords, (3) implementing a vectorizer (CountVectorizer) in order to represent each entry as a count of unique words, and (4) implementing a LDA model that outputs *k* topics for column *c* in the data. I fit the model to the cleaned survey results, represented as a single PySpark DataFrame, and outputted the results as a table of sets.

I spent some time adjusting the number of topics. While I wanted to maintain low log perplexity and high log likelihood (my two evaluation metrics), I also wanted to maximize (subjective) intelligibility for a human reader. I settled on k=15 as a balance between acceptable evaluation metrics and high coherence for a human reader.

2. **How do the descriptions of the aftermath of incidents differ between professor/faculty/department chair perpetrators and non-faculty perpetrators?**

I noticed the prevalence of incidents with professor/faculty/chair perpetrators while reading through the survey results. However, this distinction was not captured in the survey in any systematic way (again, user-determined text fields!) I decided to generate this categorical variable myself by searcing for "prof", "faculty", or "chair" substrings in the "perpetrator" field (I first looked just for "professor", but noticed this excluded related data.)

I decided to characterize the "descriptions" not as a topic modeling question, but as an n-gram question. That is, what are the common sequences of words of *n* length that appear in descriptions of professor/faculty/chair incidents versus non-faculty incidents? I was curious how these results would differ from #1.

To categorize entries as professor and non-professor, I converted the PySpark DataFrame to pandas and added a new column, "IsProf", which contained a boolean value indicating if the substrings were found. I then converted back to a PySpark dataframe (it would have likely been possible to do this within PySpark, but I couldn't figure out how!) I now had two distinct PySpark DataFrames to work with: professors_spsark and no_professors_spark.

With the subsets assembled, I approached this problem using a PySpark MLLib Ngram pipeline. Beginning with the chosen subset of data (professor/no professor), I tokenized responses for a given column, removed stop words, and generated NGrams of length *n*. I then transformed the output back into a pandas DataFrame to output the n-gram names and frequencies. I tried out several different combinations of n-gram lengths to characterize the aftermath of events (i.e outcome and impact fields, not "event"). I visualized the outputs as seaborn barplots.

3. **What language in incident descriptions is most predictive of whether the perpetrator was a professor/faculty/department chair or not? How powerful is this prediction?**

For this task, I took inspiration from the ranking of feature importance in one of our homework assignments. I thought it would be helpful to build a classification model that predicts professor or no professor for the perpetrator. My main interest was to see to what degree I could generate a realistically predictive model for this category based on the incident descriptions. If so, what specific fields of description (such as the event, the punishment, etc.) most strongly predicted professor/no-professor perpetrators? And given those fields, what features (in this case, words within the description) are the strongest predictors within the model?

To start answering these questions, I went back to the IsProf categorical variable generated in pandas. I converted this boolean value to a 1/0 integer encoding to use for classification later on. Instead of generating two subsets of data, I joined this additional category to all of my data in the original dataset. This resulted in a transformed PySpark dataset containing the full survey responses along with a new IsProf categorical variable.

I approached this question as a machine learning/classification task, and built a classifier pipeline using PySpark with a Random Forest classifier. To do this, I wrote a function that establishes a new pipeline for classification. I used some elements of the LDA pipeline earlier, such as starting with stop words removal and tokenization. As with LDA, I included a CountVectorizer vectorizer in order to represent entries in a given column as a collection of word counts, given a specific size of dictionary and minimum document frequency for terms. These counts act as features that I then passed into a Random Forest classifier (which is, technically, a meta-classifier of tree classifiers). Finally, I set the one-hot "IsProf" variable as the target label for classification.

This CountVectorizer + Random Forest approach felt like a clear way to pursue classification without further transforming the feature set (for example, I thought about but eventually decided against word embeddings, as I wanted to make sure the resulting features would still be human-readable – in this case, a ranked list of words.

I trained and tested the model on a 80/20 split of the input data (after first forgetting to create a training set! I describe this in the results section.) I outputted accuracy as a percentage of successful predictions, and finally outputted a list of features ranked by importance in the model. I printed the success rate to screen and visualized the features ranked by importance as a seaborn barplot.

4. **Finally, what language in the incident report is most predictive of whether the respondent "left" or "quit" their position/academia/the discipline/etc.? How powerful is this prediction (in general and relative to the professor model)?**

The approach to question #4 mirrors #3 very closely. This time, I wanted to set as the predicted variable whether the respondent described leaving or quitting their position/academia/the discipline/etc after the event. I searched for "left" and "quit" substrings in any of the personal outcome categories (life, mental, career) and encoded the result as a new one-hot variable, which I joined to the DataFrame of all survey responses.

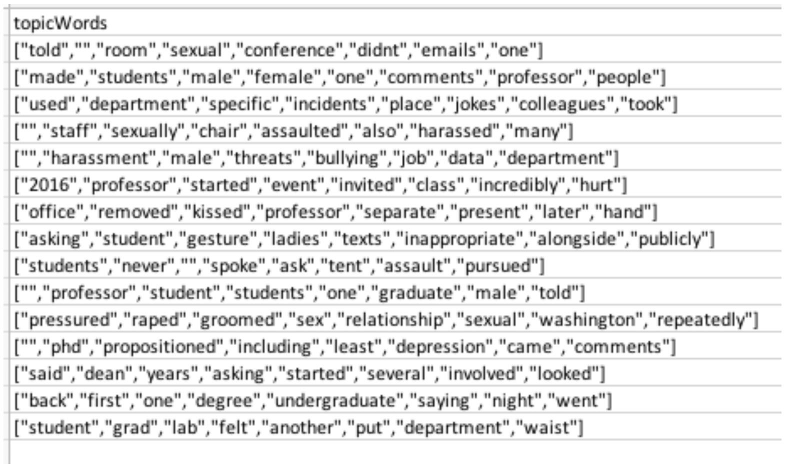
As with #3, I built a classification function based on using description words within a given column as predictors of the "left/quit" outcome. I created a Random Forest pipeline using the same sequence of tokenization, stop word removal, transforming column-specific descriptions into Count Vectors, and running this through a Random Forest classifier with Left as the label column. After fiddling with parameters, I decided to implement an 85/15 split this time. Finally, I fit this model on the data, transformed my predictions, outputted success rate, and outputted and then visualized the features sorted by importance in the model.

4. Analysis and Results

1. **What are the main themes (or topics) in reports of harassment/assault event?**

A number of clear themes emerged in the descriptions of sexual harassment and assault on campus. The topic modeling approach succeeded in describing the domain of possible descriptions (total set of topics) for specific description fields, as well as showing how language for a given topic is related (associated words within a topic. I believe these findings may correspond to real patterns in incidents, outcomes, and impacts worth investigating further.

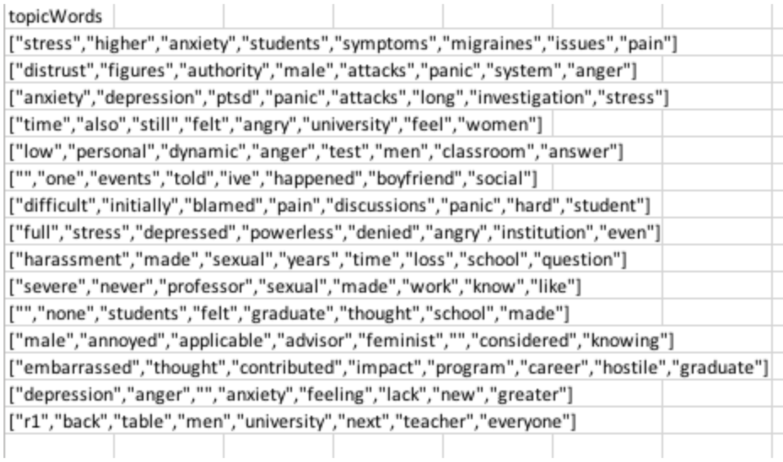
To illustrate patterns in the **event,** here is a table of topics in the **event description category:**



topicWords
["told","","room","sexual","conference","didnt","emails","one"]
["made","students","male","female","one","comments","professor","people"]
["used","department","specific","incidents","place","jokes","colleagues","took"]
["","staff","sexually","chair","assaulted","also","harassed","many"]
["","harassment","male","threats","bullying","job","data","department"]
["2016","professor","started","event","invited","class","incredibly","hurt"]
["office","removed","kissed","professor","separate","present","later","hand"]
["asking","student","gesture","ladies","texts","inappropriate","alongside","publicly"]
["students","never","","spoke","ask","tent","assault","pursued"]
["","professor","student","students","one","graduate","male","told"]
["pressured","raped","groomed","sex","relationship","sexual","washington","repeatedly"]
["","phd","propositioned","including","least","depression","came","comments"]
["said","dean","years","asking","started","several","involved","looked"]
["back","first","one","degree","undergraduate","saying","night","went"]
["student","grad","lab","felt","another","put","department","waist"]

These topics provoke a number of questions. Is there an association between the “dean” position and long-term, sustained harassment (“years”, “several”, “involved”?) When events involve another student, why does “lab” show up frequently? Does “removed” point to a common strategy of shuffling around perpetrators without terminating employment? Also, the relationship between “pressured, “raped,” and “groomed” seems to tell a specific story.

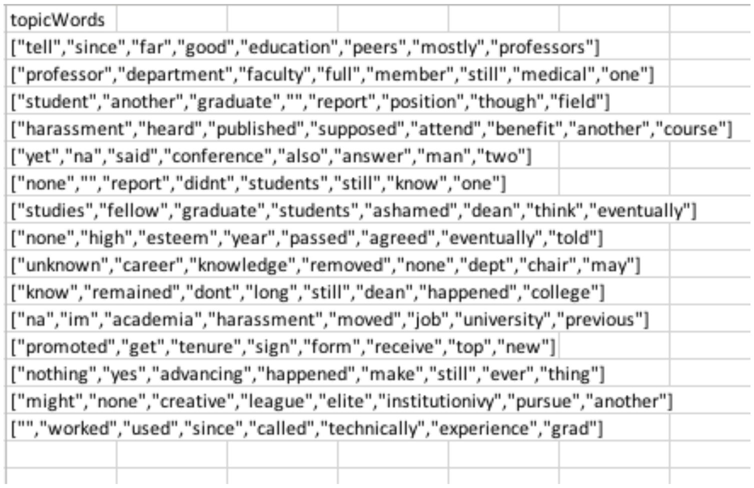
To illustrate patterns in **impact on the target**, here is a table of topics in the **mental category:**



topicWords
["stress","higher","anxiety","students","symptoms","migraines","issues","pain"]
["distrust","figures","authority","male","attacks","panic","system","anger"]
["anxiety","depression","ptsd","panic","attacks","long","investigation","stress"]
["time","also","still","felt","angry","university","feel","women"]
["low","personal","dynamic","anger","test","men","classroom","answer"]
["","one","events","told","ive","happened","boyfriend","social"]
["difficult","initially","blamed","pain","discussions","panic","hard","student"]
["full","stress","depressed","powerless","denied","angry","institution","even"]
["harassment","made","sexual","years","time","loss","school","question"]
["severe","never","professor","sexual","made","work","know","like"]
["","none","students","felt","graduate","thought","school","made"]
["male","annoyed","applicable","advisor","feminist","","considered","knowing"]
["embarrassed","thought","contributed","impact","program","career","hostile","graduate"]
["depression","anger","","anxiety","feeling","lack","new","greater"]
["r1","back","table","men","university","next","teacher","everyone"]

These topics clearly show how devastating the effects of sexual harassment and assault can be. Impacts raise from higher stress and anxiety, distrust of men and authority figures, depression + ptsd, all the way to embarrassment and hostility, unhealthy dynamics, and powerlessness and being denied within an institution.

To illustrate patterns in the **outcome for the perpetrator,** here is a table of topics in the **punishment category:**



topicWords
["tell","since","far","good","education","peers","mostly","professors"]
["professor","department","faculty","full","member","still","medical","one"]
["student","another","graduate","","report","position","though","field"]
["harassment","heard","published","supposed","attend","benefit","another","course"]
["yet","na","said","conference","also","answer","man","two"]
["none","","report","didnt","students","still","know","one"]
["studies","fellow","graduate","students","ashamed","dean","think","eventually"]
["none","high","esteem","year","passed","agreed","eventually","told"]
["unknown","career","knowledge","removed","none","dept","chair","may"]
["know","remained","dont","long","still","dean","happened","college"]
["na","im","academia","harassment","moved","job","university","previous"]
["promoted","get","tenure","sign","form","receive","top","new"]
["nothing","yes","advancing","happened","make","still","ever","thing"]
["might","none","creative","league","elite","institutionivy","pursue","another"]
["","worked","used","since","called","technically","experience","grad"]

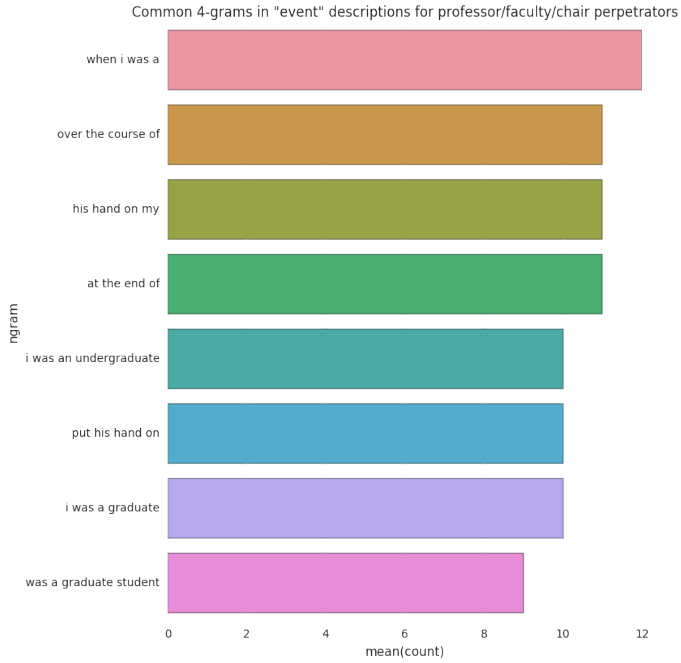
Here, we find very little evidence of a repercussion like firing, although “removed” does appear once. Instead, topics emphasize no punishment or unknown punishment, and even point to some positive outcomes such as “promoted”, “get”, “tenure”.

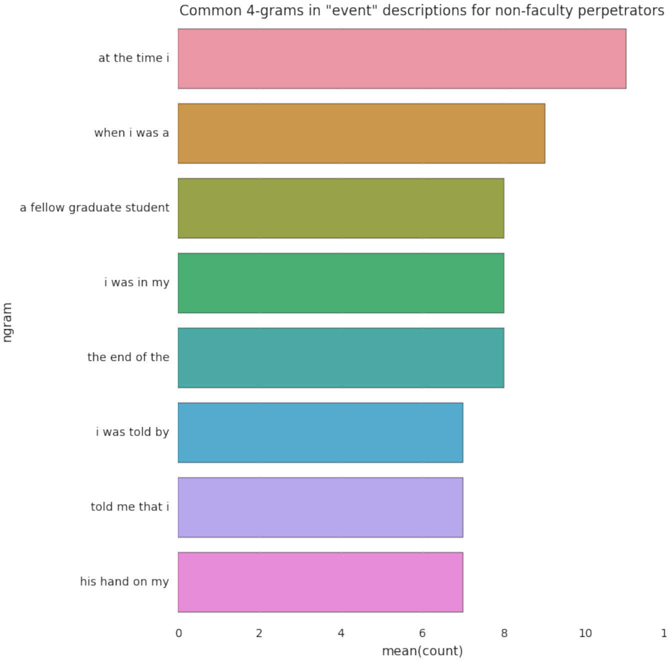
These topic results alone do not provide conclusive evidence of the average descriptions for these categories. However, they do sketch out the domain of possibilities and show relationships that tell cohesive stories about common elements in these narratives, which I hope will be helpful in determining the scope of future study. We go further into these possible trends in the following explorations.

2. **How do the descriptions of the aftermath of incidents differ between professor/faculty/department chair perpetrators and non-faculty perpetrators?**

Here, I worked under the hypothesis that professor and non-professor perpetrated incidents would vary in some meaningful way. The evidence uncovered from n-gram outputs indeed reveals a few interesting distinctions between these subsets.

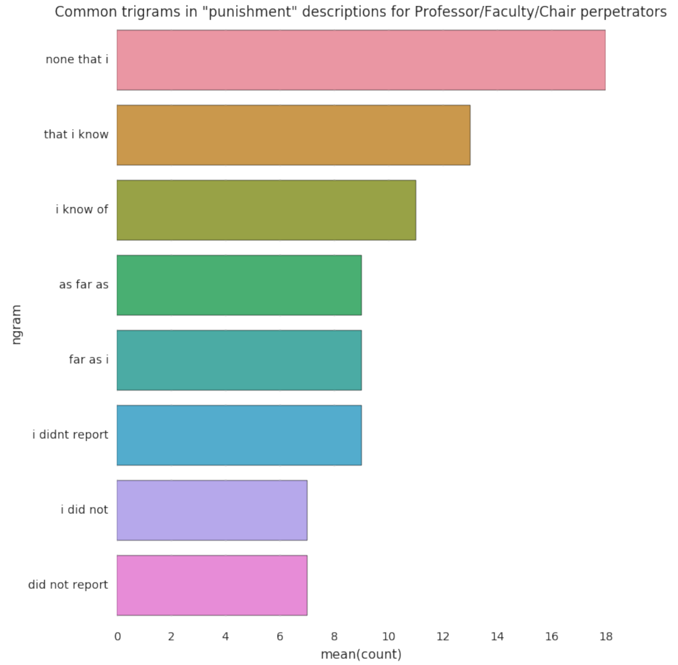
Comparing 4-grams in “event” descriptors between professor/faculty/chair perpetrators and non-faculty perpetrators:

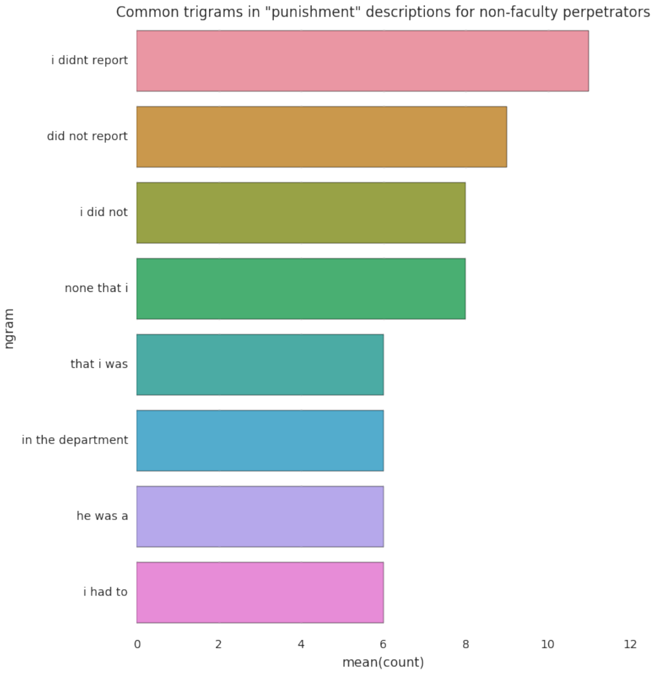




The 4-grams in both subsets share some common elements, such as use of he/him pronouns and putting "his hand on my," which signal physical contact from men. Interestingly, "over the course of" ranks very high for professor/faculty/chair perpetrators, which may imply long-term or repeated harassment/assault. Professor 4-grams also point to events during both graduate and undergraduate years. In contrast, non-professor 4-grams mention "a fellow graduate student" and commonly make reference to being "told" something.

Now, let's compare visualizations of trigrams of punishment descriptions between professor and non-professor perpetrators:





Here we find a very clear pattern with professor perpetrators: no known punishment. A secondary pattern is that the respondent did not report the incident. Interestingly, the dynamic is reversed for non-faculty perpetrators: primarily these descriptions emphasize not reporting the incident, and secondarily mention none that I… (which seems to imply no known punishment).

The n-gram outputs provide further detail about the domain of incidents. While again we cannot say conclusively, the data seems to point to sustained harassment and assault from professor/faculty/chair perpetrators towards undergraduate and graduate targets that receive no punishment. Meanwhile, non-professor perpetrators may often be graduate student peers, and may not report at a higher rate (and/or result in punishment more often).

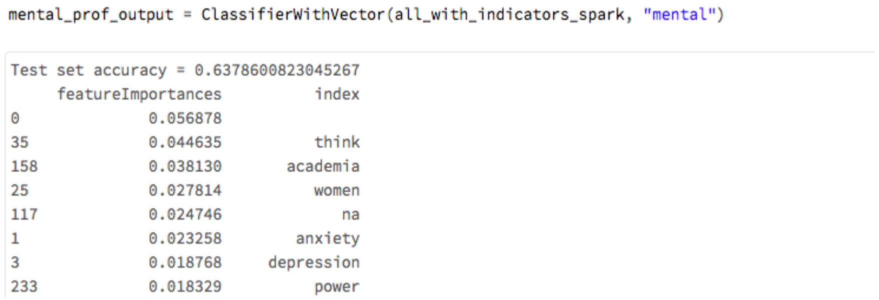
3. **What language in incident descriptions is most predictive of whether the perpetrator was a professor/faculty/department chair or not? How powerful is this prediction?**

After much tweaking, the classification model to predict professor/not professor perpetrators resulted in modest success. The most accurate prediction occurred with the “event” column – however, I chose to exclude this result given that the most accurate feature by far was “professor,” and simply mentioning a professor in both the perpetrator and event field didn’t feel meaningful in any particular way. I instead focused on the perpetrator outcome and target impact.

At first, I thought I had achieved a very strong result: accuracies in the range of .72 to .8! However, I noticed my mistake: I had failed to create a distinct test and training sets. The model had clearly been overfitted.

Once I created a test/train split, the correctly fit model ranged from .59 to .637 accuracy in predicting professor status. Even though there was a modest difference in the professor/no professor n-grams in #2 above, the classification task produced minimal results. I interpreted this as signifying that the modest changes observable in the aggregate (via n-grams) were too subtle to successfully model in this kind of classification task.

I felt that, given these very low numbers in terms of classification success, it would be misleading to try to interpret the featureImportance figures in great detail. For instance, here is the raw output of the most successful classification task (.637 accuracy) which occurred with the **mental health impacts** category:



```
mental_prof_output = ClassifierWithVector(all_with_indicators_spark, "mental")
```

```
Test set accuracy = 0.6378600823045267
     featureImportances          index
0              0.056878
35             0.044635          think
158            0.038130       academia
25             0.027814          women
117            0.024746             na
1              0.023258        anxiety
3              0.018768     depression
233            0.018329          power
```

I am not sure how to interpret this outcome. Mentioning “think”, or a blank entry (somehow this persisted even after attempting to remove nulls at length!) do not easily lead to a real-world interpretation, and “academia” and “women”, while interesting, elude easy interpretation as well.

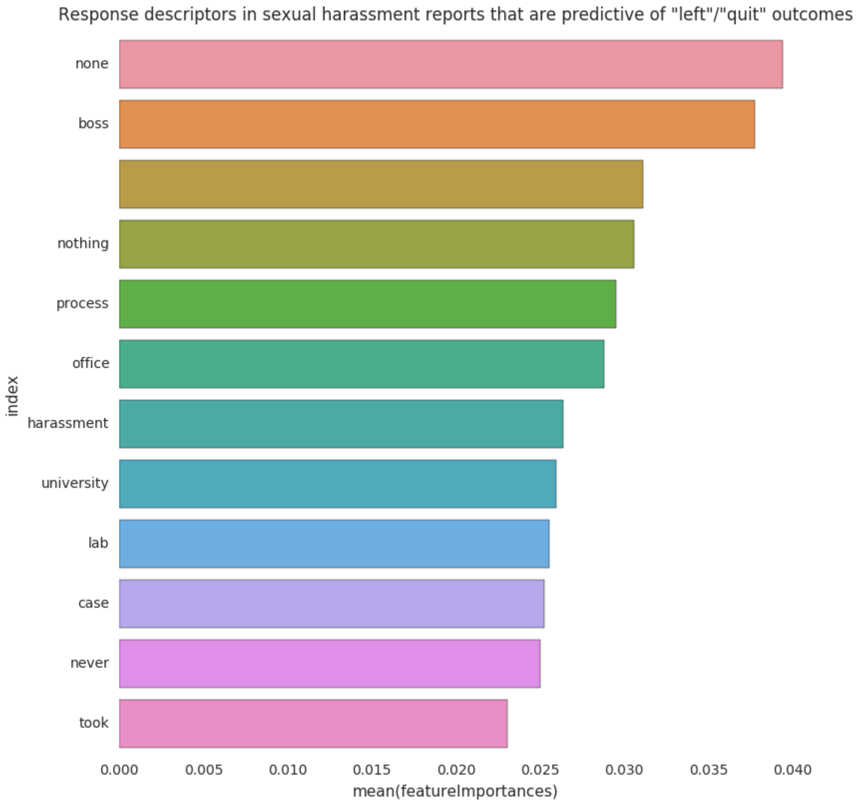
The main finding from this question is that the current descriptions only weakly predict professor/no professor. Perhaps another approach would lead to a more successful model, but at the moment, the relationship doesn’t appear to be very strong.

4. **Finally, what language in the incident report is most predictive of whether the respondent “left” or “quit” their position/academia/the discipline/etc? How powerful is this prediction (in general and relative to the professor model)?**

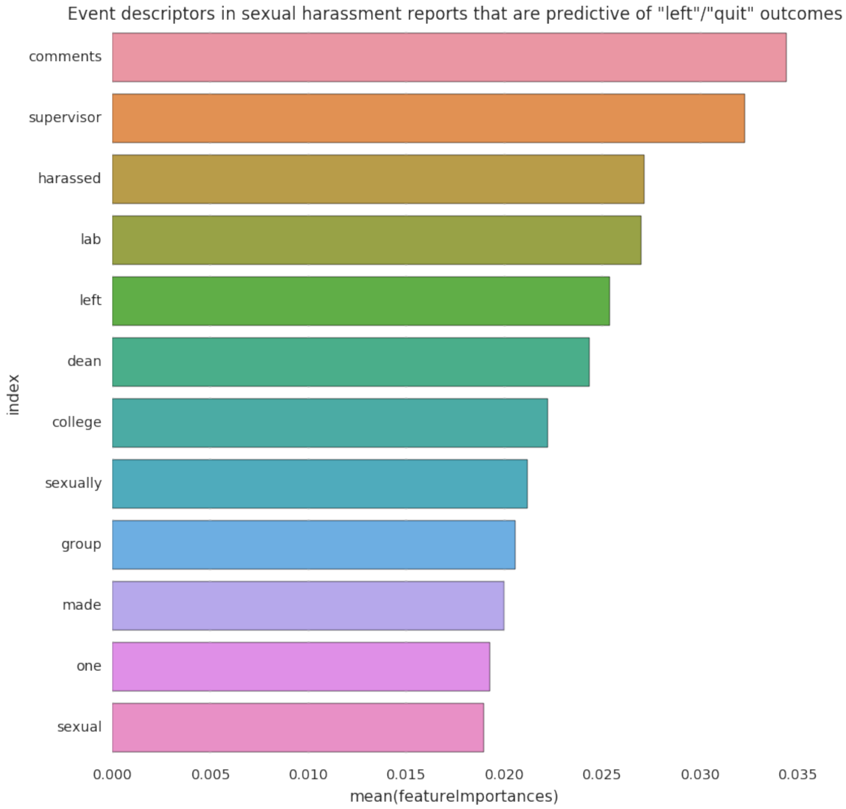
In the previous step, we achieved classification success of just .59 to .637 in the attempt to predict professor status of perpetrator. In contrast, I discovered that the task of predicting the target leaving or quitting a job/academia/etc. based on either descriptions was a much higher success rate of **79.3%!!**

This signifies much stronger evidence of predictive power in this model. In other words, it appeared the existing data lends itself much more strongly to predicting whether an individual will leave academia or quit their job after an incident as opposed to whether the perpetrator was a professor. From a domain perspective, this finding has a strong potential to tell a compelling story about how sexual harassment and assault incidents drastically alter the lives of targets.

Given the strong predictive power of this model, I felt it was appropriate to delve into the specific features and their relative importance. This resulted in what i believe to be some of the strongest findings of the project:



Analyzing the “response” descriptions in order to predict a “left”/”quit” outcome reveals that the strongest predictors appear to be a non-response (“none”, blank, “nothing”, “never”) or mention of a supervisor (“boss”). We cannot directly infer the directionality of this association in the model, but it definitely gives us pause to think about how a non-response might be encouraging individuals to abandon their career ambitions entirely.



Quite similarly, events that include references to people in direct positions of power (“supervisor”, “dean”) appear predictive of an individual leaving or quitting their position or academic career. Again we see the lab setting referenced meaningfully here. I am very curious what “comments” might mean in this context – perhaps a reference to how the incident is discussed?

These two visualizations point to a possible correlation between individuals experiencing assault and harassment from their direct supervisors and eventually leaving or quitting their career/academic pursuits, as well as a relationship between non-responses after incidents and leaving/quitting. This might compel us to direct our attention towards the direct supervision of individuals – and from previous results, it appears this individual may likely be a graduate student – as a focal area for exploring this data further, as well as increasing the response rate to sexual harassment/assault claims across the board.

5. Conclusion

My hope in working with this data is that I would learn more about iterating quickly and creatively through various NLP/machine learning methods to understand data better. I believe the outputs clarified more about areas subsequent scholarship and activism might focus on: power differentials and supervisor relationships, why targets quit their jobs or leave the field, the importance of specific settings like the lab or a conference, the possibility that professors and direct supervisors are long-term perpetrators who are sheltered by their departments and receive no direct punishment or even institutional response, etc.

I believe this work raises yet more questions, and points to the immense value of uncovering detailed data about sexual harassment and assault on campus in order to develop targeted, effective interventions that hold perpetrators accountable.