Characterization and Comparison of Topics and Top Words in the H2-A Job Postings Between TRLA Catchment States and non-TRLA Catchment States

You-Chi Liu (you-chi.liu.23@dartmouth.edu

Laura Holland (laura.holland.22@dartmouth.edu)

Josh Calianos (joshua.22@dartmouth.edu)

Abstract

The Department of Labor (DOL) conducts the H-2A visa program, which allows agricultural employers to hire foreign guest workers on temporary work visas. While the programs provide important employment opportunities for low-wage workers, those working under these visas can also be especially vulnerable to violations of their legal rights. Using comprehensive datasets of job postings from Texas Rio Grande Legal Aid, we implemented topic modeling to explore the top topics and words within job postings for the H-2A visa program. The major variation we explored was between topics and words in job postings from the TRLA-catchment area versus the top words and topics in states not represented by TRLA. Using TF-IDF scoring, we found that the TF-IDF scores were generally higher for top words in non-TRLA corpus, however, direct comparison was difficult due to the difference in the number of postings between TRLA-states and non-TRLA states. Few top words found in both corpuses were distinct or related to things other than hiring processes.

1 Introduction

In recent years, the Department of Labor (DOL) has seen a rapid increase in the popularity of their H-2A program [2]. This program, which aggregates farm jobs and provides temporary visas for migrant farmworkers, has quadrupled in size since 2010 [4]. In fiscal year 2019, employers posted over 250,000 jobs, for which they provided over 200,000 visas [3]. Similarly, albeit on a smaller scale, the H2-B visa program is aimed at seasonal, non-agricultural workers.

To participate in these programs, American employers must certify that they are facing a shortage of American workers and that they are upholding health and safety conditions [1]. Employers who violate these workplace safety rules are subject to fines and penalties, although the rate of penalization is assumed to be far below the true rate of rule violations [1]. In addition, employers who have been sanctioned can initiate a long process of appeal and arbitration.

The decentralization of the H-2A and H2-B programs exacerbates these problems of punishing violators. Many federal agencies are responsible for administering the H-2A and H-2B program requirements, including the DOL, the Department of Homeland Security, the Department of State, the Department of Justice, and the Department of Health and Human Services (HHS). Given how many different agencies are responsible for monitoring violations, data regarding employers and violations is incredibly disorganized.

While the programs provide important employment, housing, and visa opportunities for workers, these workers can also be especially vulnerable to violations of their legal rights. There is no cohesive enforcement across agencies and reductions to enforcement budgets have weakened oversight mechanisms. While employers sign a contract with the DOL promising to abide by worker protection laws, investigations show that employees still face systematic abuses [2]. Compiling datasets of alleged violations, across agencies and across time periods, is a large step forward in combating future abuses.

The Texas Rio Grande Legal Aid (TRLA) has responded to advocates across the country to create the most comprehensive and up-to-date database of H-2A and H-2B job postings. TRLA is the second-largest civil legal aid organization in the United States. The agricultural labor law practice within their broader labor and employment practice group represents farmers from six states: Texas, Kentucky, Alabama, Mississippi, Louisiana, Georgia, and Tennessee.

This paper aims to build upon a separate dataset that TRLA has collected, namely one of job addendums that go above and beyond boilerplate H-2A job postings. These addendums provide one of the largest original text datasets for H-2A postings in the country. Accordingly, this paper explored the

textual data provided with the goal of modeling and identifying common themes and topics across the postings.

Topic modeling provides us with methods to organize, understand, and summarize large collections of textual information. Topic modeling is a type of statistical modeling for discovering the abstract "topics" that occur in a collection of documents of words [6]. In this report, our aim is to use topic modeling to analyze the topics and top words in job postings from the TRLA catchment states (TX, MS, LA, KY, AL, TN) and compare them to postings in other states. Although our results were fairly banal, we believe that they will serve as an important first step in a more detailed analysis of H-2A violations in the future.

2 Related work

Our work is not the first to examine H-2A job postings. A 2015 Government Accountability Office report examined the H-2A and H-2B programs with the aim of increasing protections for foreign workers [2]. They found that a common "abuse" was providing insufficient job information [2]. Although this report did not examine job addendums, examining addendums can shed light on many of the problems that the GAO discovered. For example, the lack of information on many jobs provides a greater opening for illegal job recruiters to act as a middle-man between employers and workers. These illegal recruiters frequently traffic in information about a job and require a (similarly illegal) payment to connect a job-seeker to an employer [2]. These payments may place workers in debt to the recruiters, which has been shown to increase the probability of gender-based discrimination in future jobs and future human trafficking by the recruiter [5]. Other frequent violations included those related to pay and the safety of employee housing [2].

As described above, the federal government has little capacity to investigate these violations. A 2020 report from the Economic Policy Institute detailed how the Department of Labor's Wage and Hour Division (WHD) had merely 1,500 employees in 2019 although they oversaw over 148,000,000 workers [7]. This report detailed how agricultural employers have disproportionate violations relative to their number of employers. However, only 1% of employers are investigated each year (p. 5-6)

The literature is at a point where methods to optimize the WHD's violation-finding process would be well received. Because past violations by an employer are poor predictors of their future violations, other methods are required [7]. We hope to add to that conversation with our analysis of the job addendums, which we believe to be currently under investigated by the literature.

3 Data

The first dataset that we used for this project was obtained by TRLA through a Freedom of Information Act (FOIA) request to the Department of Labor. Originally, the FOIA data acquired was in PDF form. Before we received the data, the PDFs were parsed in a process that retained the job number associated with the PDF, the section for which the addendum was added, and the addendum text itself. The data comprised five fiscal quarters, from January 2020 to March 2021.

The original unit of analysis in the FOIA data was sections of each job. This meant that job postings with multiple addendums could take up multiple rows. The first step in our analysis was aggregating these rows with identical cases but different sections into a single unit. We hoped that all the job listings would be in English; however, we later figured out that some of the listings are in Spanish. This made proceeding directly to text analysis difficult.

Without further data cleaning, the mix of two languages would make our topic modeling less accurate. We were hesitant about directly deleting Spanish job listings as well because many listings included English and Spanish in the same cell without any structured order of the languages. We worried about omitting important correlations and findings. Ideally, we wanted the ability to directly relate frequent keywords to its section. This would help us further our comparisons and analyses of the importance of the words in TRLA and non-TRLA catchment areas. Discussion of our answers to these questions is in the "Methods" section below.

The second dataset that we used for this project was the H-2A Programs' Disclosure Data. We acquired this public data from the DOL's website. To be consistent, we decided to only use data that also was in the same time window as the first data – Q1 2020 to Q1 2021. Accordingly, we began with two Disclosure Files for both 2020 and 2021. The original units of analysis for this data were the unique H-2A case numbers, which aligned to the unit of analysis that we are using. This meant that each row was intended to be a unique job posting, although some job postings sought dozens or even hundreds of employees for identical roles. Unexpectedly, there were some duplicate rows because some employers applied for H-2A status multiple times for the same job. We deduplicated our data accordingly. We kept the first posting in case we wanted to analyze seasonal trends in the postings. Although we did not go down that path, future researchers may wish to use the last listing.

A handful of job listings (sixteen out of over ten thousand) did not have a state listed for the job location, but some of them did have a job listed in the "POC State" category. We imputed this "POC State" data when there was no original state listed.

After our deduplication, we continued to use the case number as the unit of analysis. Wage offer was an important index to look into especially in the case of job listings. One limitation was a

category of salaries ("WAGE_OFFER") which could be expressed in per month, per hour, aond per year figures. Additional steps were needed to make the unit consistent in order to compare this variable between TRLA and non-TRLA listing and across listings that were removed, in Spanish, or kept. Additionally, different jobs required different working hours. The hours per for each job varied significantly throughout employers, so we resorted to adding the expected hours for each of the seven days of the week into an additional column, which we called "Anticipated Number of Hours" Although this did not directly affect our analysis in this paper, we discuss below how this calculation may be good practice for future uses of this dataset.

4 Methods

After deduplicating the rows and mutating some columns, we cleaned the dataset further and tested some of the assumptions that any causal result from this dataset would have to rely upon.

First, we decided to remove job addendums with Spanish-language text. These postings made up fewer than 9% of all addendums, and including the Spanish risked introducing measurement error during the text analysis phase. We found that Spanish-language jobs were more likely to be in the TRLA catchment zone, which was expected because of the large proportion of Spanish-speaking migrant workers who work in Texas. We did not find any other significant differences between English and Spanish postings.

Second, we tested that our analysis of jobs with addendums would have validity extending to all of the H-2A jobs. Only a small fraction of the jobs in the full H-2A dataset were matched with job addendums. To ensure that this was a representative subset, we joined the data in two ways. First, we left-joined the full job dataset to the job addendum data by the case number column. After subsetting that merged dataset to the job listings without addendums, we concluded that these jobs were not significantly different from the jobs with addendums after testing some of the quantitative variables. We found only small differences between the median values of these two types of postings; for example, the median number of workers sought in jobs with addendums was 6 instead of 4. See Table 3 for this balance test.

After this cleaning, we kept the left-joined dataset to analyze as our primary dataset. We assigned an indicator variable to rows that were in the TRLA catchment region and found that they were a smaller proportion of posts with addendums than posts without addendums. We did not find this concerning – much of the variation between the two groups simply came from the fact that non-TRLA jobs were much more likely to have addendums concerning transportation to and from the job, because those jobs are farther from a border.

With this cleaned data, we began our text analysis). We created a list of stopwords, starting with the default stopwords from the NLTK package [8]. We added all words with the string "work" as well as a handful of other words to try and isolate language that was as far from boilerplate as possible.

We initiated the Porter Stemmer from the NLTK package to stem the words to their roots, then mapped the relative frequencies of words in the addendums and compared the most frequent words in the TRLA states as well as those outside [9].

We utilized Term Frequency - Inverse Document Frequency (TF-IDF) scores weighting to extract the corpus-level top words in our raw data which included the entire corpus, TRLA states, and non TRLA states. TD-IDF scores measure how relevant a word is to a specific topic in a document. The score is compiling using a combination of how many times the word appears in the document and the inverse document frequency of a word across a set of documents.

We then used Latent Dirichelet Allocation (LDA) modeling to run topic modeling on the entire corpus, TRLA states, and non TRLA states. We identified the high-probability words for each topic. The LDA topic modeling method does not categorize topics (instead producing topic_0, topic_1, etc.), so we then categorized each topic based on the high-probability words within the topics.

The last step was producing visualizations that were helpful for viewing the results of the TD-IDF scores and the LDA modeling. Visualizing topic models is difficult within pandas, but we were able to produce bar charts for both the TD-IDF scores and the LDA topic modeling.

5 Results

Corpus Descriptors

The top five sections of the addendum are shown in Table 1 below. They were "Job Requirement," "Daily Transportation," "Pay Deduction," "Inbound/Outbound Transportation" and "Job Duties."

SECTION_ NAME	SECTION_ NUMBER	SECTION_DETAILS
Job Requirements	B.6	Three (3) months experience with references required as a beekeeper. May not have bee, pollen or honey-related allergies. Must be able to obtain a drivers license within 30 days following hire and obtain clean driving record. Must be able to lift 75 lbs. Once hired, worker may be required to take a random drug test at no cost to the worker. Testing positive or failure to comply may result in immediate termination from employment.
Daily Transportation	F.1	Living & laundry facilities available. Housing will be clean and in compliance with OSHA Housing Standards at 29 CFR 1910.142 when occupied. Workers will be responsible for maintaining housing in a neat, clean manner. Housing and utilities are provided at no cost to workers who are unable to return to their place of residence the same day.
Pay Deductions	A.11	Reasonable repair cost of damage from deliberate destruction, other than that caused by normal wear and tear, will be deducted from the earnings of workers found to have been responsible for damage to housing, furnishings, or equipment. No deductions will be made which would bring the employee's hourly wage below the Federal Minimum Wage. (Reference; Internal Revenue Service Publication 51, Circular A, Agricultural Employer's Tax Guide). Garnishments and levies required by law shall be deducted...
Inbound/Outbound Transportation	F.2	Alternatively, the employer may, at his or her discretion, provide transportation funds or purchase transportation tickets for the worker ahead of time. Upon completion of the work contract or termination of the worker's employment without cause, the employer will cover the reasonable expenses for the worker's return trip back to the place from which the worker came to work for the employer. Transportation payment will be equal to the most economical and reasonable common carrier transpo...
Job Duties	A.8a	Workers may not entertain guests in employer-provided housing premises after 10:30 PM, except on Saturdays when guest hours end at 12:00 midnight. No persons, other than workers assigned by employer, may sleep in housing. Workers may not deliberately restrict production or damage products/ commodities. Workers may not physically threaten other workers, the employer, supervisors, or members of the public with any tool or weapon. Workers who violate this rule may be subject to ...

Table 1: **Top 5 sections that contain the most listings in the original uncombined job addendum.** Section Name shows the name of the section that listings were initially categorized. Each section had its own number and section details column showed examples of listings in the corresponding section.

After merging and cleaning data, Table 2 allowed us to visualize the number of cases. We counted 13,127 cases and listings over five fiscal quarters; however, it was important to note that the number of original cases was higher than this number. This analysis was conducted after replacing some missing values in the Employer State column, so we could infer that the Employer State column has more missing values that Employer POC [Place of Contact] State column.

	CASE_NUMBER	JOB_DESCRIPTION	EMPLOYER_STATE	EMPLOYER_POC_STATE
count	13127	13127	13116	13116
unique	13127	9845	50	50

Table 2: **Count of the number of cases, listings, states (including both POC and non POC) after inner joining the disclosure and addendum data together by case number**.

This figure shows the number of combined listings used in topic modeling and text analysis. The count of the number of EMPLOYER STATE after replacing the missing value with POC State was crucial because it could allow us to understand the proportion of the states that were in TRLA and TRLA catchment areas. Since there were 50 unique count in both EMPLOYER STATE and POC State columns, we knew that the merged data did encompass listings from all the States; however, the exact distribution needs further examination.

Our major comparison regarded jobs within the TRLA catchment area versus those in remaining states. After merging, the total number of job orders was 13127 (which included 3684 TRLA and 9443 non-TRLA catchment areas) for five fiscal quarters; however, it was important to note that the number of job orders was higher than this number. According to figure XX, the proportion of TRLA catchment areas' job listings in non-English addendums is 30.6 percent which occurred to be higher than those in English addendums; the difference was around 7 percent. This result showed that there are proportionally more non-English addendums in catchment states than in English addendums, which is indicative of the greater number of more Spanish-speaking H-2A workers in TRLA catchment states. Alternatively, the employers who were in TRLA catchment states were more willing to hire people who could not speak English.

	WAGE_OFFER MEDIAN	TOTAL_WORKERS_NEEDED MEDIAN	ANTICIPATED_NUMBER_OF_HOURS MEDIAN	PIECE_RATE_OFFER MEDIAN	TOTAL_OCCUPANCY MEDIAN	LIFTING_AMOUNT MEDIAN
English addendum	13.48	6.0	40.0	0.0	8.0	60.0
No English addendum	13.13	4.0	40.0	0.0	6.0	6.0

Table 3: **Median value of the key variables in English addendum and Non English addendum.** This table compares the median values of quantitative variables between the English and non English addendum. The values were relatively similar. One small difference we noted was that median wage was slightly lower in Non English addendum.

As we were solely looking into the English addendum and disregarding the non-English addendum, it was crucial to ensure that the key variables in the two categories of listings were not diverging significantly. From table 3, the median wage was slightly lower in the addendum that was not English, which could potentially infer the reason behind why these addendums were not listed in English and explain the comparatively low median value of total occupancy. However, the overall difference of values in each variable between the English addendum and the Non-English addendum was minimal.

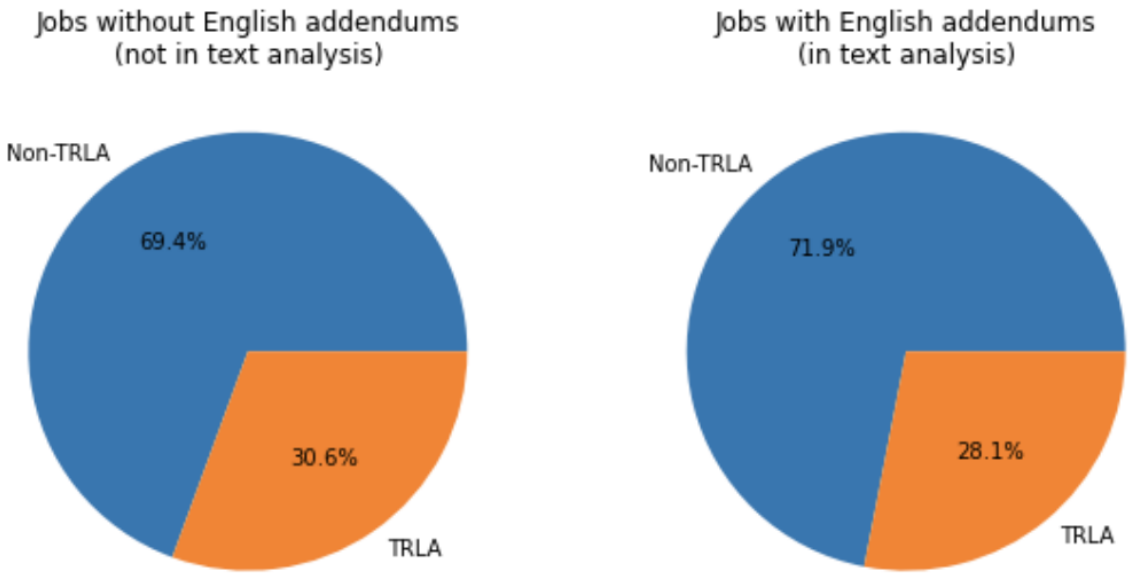

Figure 1: **The Percentage of non-TRLA and TRLA Listings in Jobs Without English Addendum sand in Jobs With English Addendums** .This figure compares the proportion of non-TRLA and TRLA job listings in the addendums that were used in the text analysis and the proportion in the dataset before cleaning out spanish listings.

Top Words Resulting From High TF-IDF Scores

Comparing Figure 2 and Figure 3, we realized that the top 10 words ranked by the TF-IDF scoring found in each corpus did overlap. Unsurprisingly, both sets centered around the same theme: jobs and hiring. Most of the top words found in the corpus were expected, as they related to jobs and hiring. If you look at Figure 2, you can see that "transportation," "outbound," and "inbound" were the top words in the non-TRLA corpus. These words make sense because according to Table 1, "Inbound/Outbound Transportation" and "Daily Transportation" are in the top 5 sections of the addendum, which means that a lot of the listings should be related to these topics. As discussed earlier, these listings are likely a figment of geography.

One word among the non-TRLA corpus that caught our eye was "termin." The word "termin" could be the stemmed result of the word "terminate." This particular word in Figure 3 caught our attention because it could potentially be related to violations, as employers could terminate the contract of their workers without a clear or legitimate reason. Digging further into the original job listings, we found that in most of the listings where the word "terminate" occurs, the conditions for termina-

tion were clearly listed. Below is a common excerpt from multiple listings that include the word "terminate."

"Employer may terminate a worker if a worker: refuses without justified cause to perform work for which the worker was recruited and hired; or commits a serious act of misconduct; or fails, after completing any training or break-in period, to be able to perform all of the tasks described in the job order. If the worker voluntarily abandons employment before the end of the contract period or is terminated for cause, and the employer notifies the SWA, DOL, and USCIS in the case of an H2A worker, the employer will not be responsible for providing or paying for the subsequent transportation and subsistence expenses of that worker, and the worker is not entitled to the three-fourths guarantee."

The frequent occurrence of this passage could suggest that it is standard legalese in non-TRLA catchment states due to prior violations by employers to terminate a worker's contract without basis or proper explanation. Another possible reason that the word "terminate" occurs in such so frequently across non-TRLA listings is that workers in some of those states have more rights than in TRLA states. For example, all of the TRLA states are "right-to-work" states. Although H-2A workers are unlikely to join a union in any state, right-to-work laws could be a reasonable indicator of poor worker protections in TRLA states. In states with stronger worker protections, employers may have to list out the behaviors which could lead to termination.

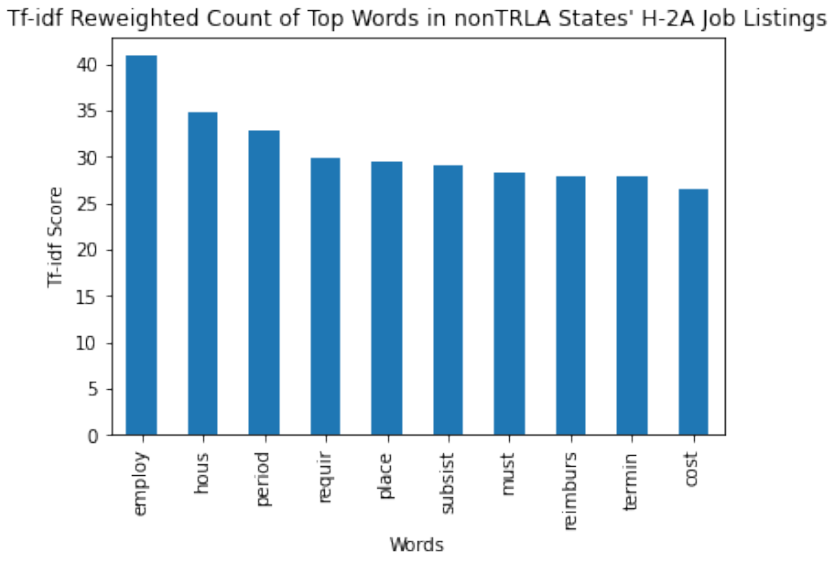

Figure 2: **Top 10 Words Ranked By TF-IDF Score in non-TRLA Corpus**

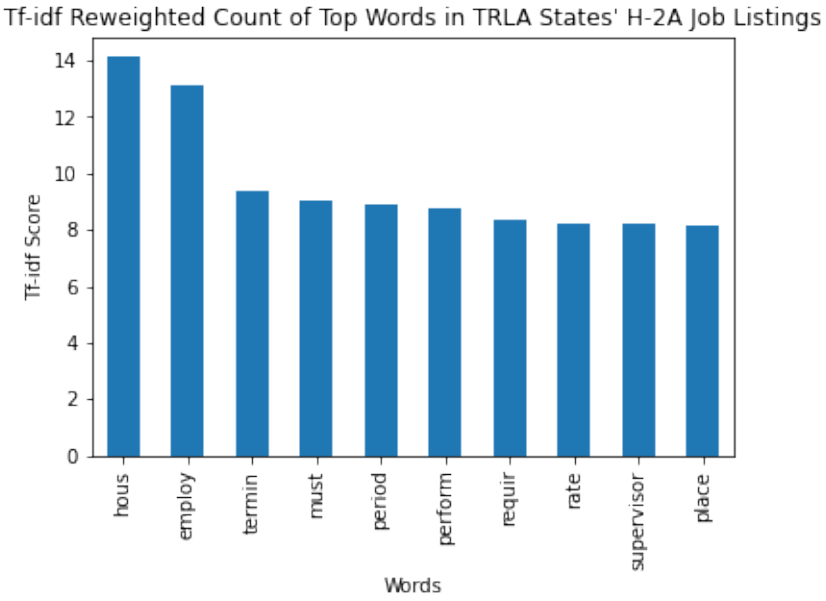

Figure 3: **Top 10 Words Ranked By TF-IDF Score in TRLA Corpus**

Topic Modeling By Linear Discriminant Analysis (LDA) Method

Figure 4 and Figure 5 presented the three topics and the associated top words that resulted from the LDA topic modeling method. The actual topics' names were not demonstrated in the output. We labeled and characterized each of the topics according to the general theme of the top words that were included in the output. The three topics in the TRLA catchment areas' job listings are "tobacco," "termination" and "compensation." Example passages of each topic are listed in Table 4. The topics "termination" and "compensation" seemed expected due to the nature of the syntax of job listings. We believe that possible reasons behind "tobacco" being a topic could be that the TRLA catchment states were more suitable to growing tobacco or had established tobacco farms. Whereas in the non-TRLA catchment areas, the job listings were harvest-related but had no specification of the type of harvest. It was interesting to note that according to Figure 5, the top words in the two topics found were nearly identical. The word "harvest" seemed to be a more identifying word, so after analyzing the job listings further, we recognize that one of the topics could be related to "harvest tools," outlining the requirements necessary to perform the job, while the other topic might be more geared toward "harvest duties." A clear example distinguishing the difference between these two topics is listed in Table 5.

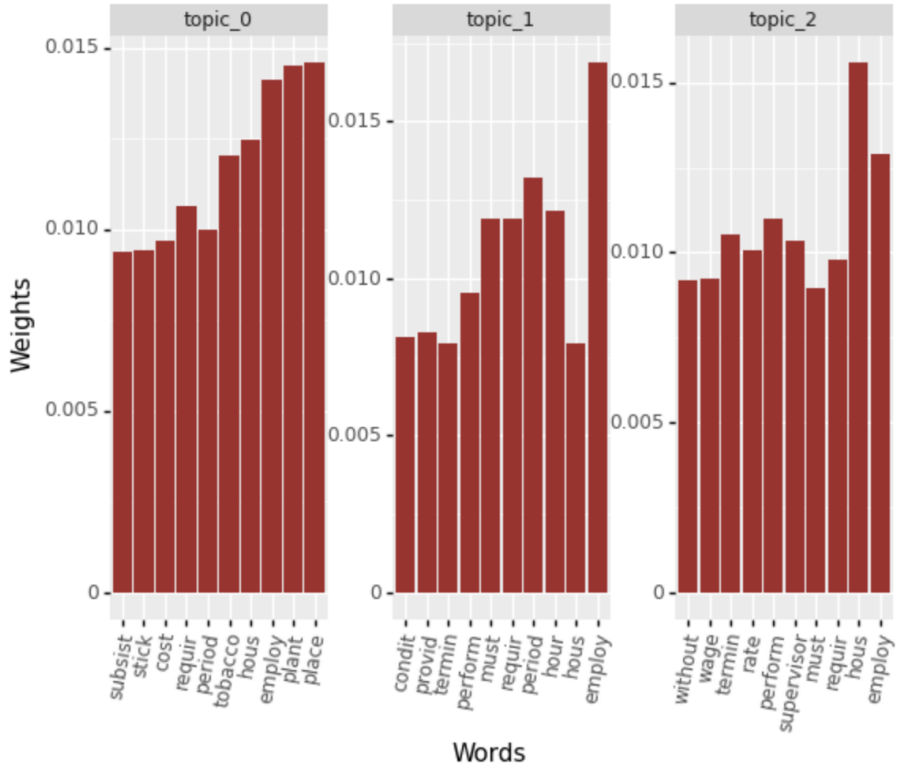

Figure 4: **Topics Corresponded Words in TRLA States' H-2A Job Listings**. Topic 0 was later labelled as the topic "tobacco." The other two topics are related to job logistics based on the top words, which we elaborate on in Table 4

Topic (manual label)	**Top words**	**Example passages from documents with high probability of topic**
Tobacco	tobacco, plant, period, subsist, stick, requir	"Workers will be trained for period of two (2) days (14 hours) after which workers will be expected to perform job required, i.e. cut 100 sticks per hour. Care must be exercised at all times to prevent bruising or breaking crops. Care must also be exercised in using tobacco knives and spears, while climbing and standing on barn rails, or with any use of equipment.
Termination	condit, hour, termin	"Employer may require post hire, random, upon suspicion or post accident drug testing, all at no cost to employee. Testing positive or failure to comply may result in immediate termination from employment."
Compensation	wage, rate, perform, supervisor, requir	"In the event that the applicable H-2A wage rate decreases for any reason during the employer?s positive recruitment or H-2A contract period in the instant job order, the employer reserves the right to decrease its offered/paid hourly wage to the new, lower wage rate, as long as the new lower rate rate remains the highest of the AEWR, the prevailing hourly wage or piece rate, an agreed upon collective bargaining wage, and the federal and state minimum wages in effect at the time work subject to the provisions of this job order is performed."

Table 4: **Manually labelled topic based on the result in figure 4**

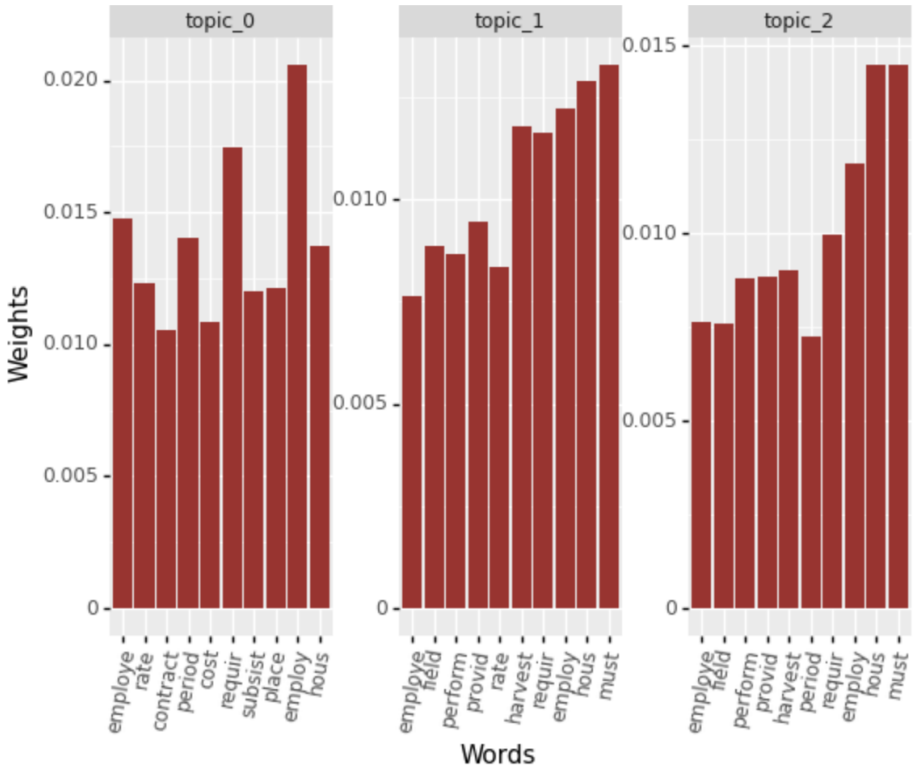

Figure 5: **Topics Corresponded Words in non-TRLA States' H-2A Job Listings**.Topic 1 and Topic 2 (later characterized in Table 5 as "Harvest Tools" and "Harvest Duties") have nearly identical top words. Topic 0 was later labelled as the topic "contracts."

Topic (manual label)	**Top words**	**Example passages from documents with high probability of topic**
Contract	contract, rate, place	"Due to the need to renew Worker's Compensation insurance each and every year, the policy may expire within the contract period requested. If the employer's Worker's Compensation policy should expire during the certified contract period, the employer agrees to renew the policy on or before the expiration date and maintain Worker's Compensation coverage for H-2A employees, and employees in corresponding employment, throughout the certified contract period.."
Harvest Tools	harvest, perform, field	"The necessary tools which could be a post type stake driver, 3 to four pound hammer or a pneumatic air hammer used in the original staking operation and tying twine will be provided by the farm.Harvest Dumper: Harvest dumper is required to stand on top of; or on the side of, harvesting containers and receive full harvest buckets that are being tossed to the harvest dumper from the harvesting employees. Bucket weights vary, based on commodity, but should not exceed 35 lbs. when filled with product and will be dumped into various types of harvesting containers."
Harvest Duties	harvest, perform, field	"Workers will be asked to alternate walking behind the transplanter and planting slips by hand using a stick. Ride a potato digger or harvester and pick up potatoes off of a chain, sort them, and place them into boxes on the digger while harvesting. Harvesting may also be done by hand in 40 lb. buckets. Clean, pack and load harvested products; place produce into coolers."

Table 5: **"Manually labeled topic based on the result in figure 5"**

6 Discussion and Conclusion

As discussed above, our biggest limitation was our sole usage of the LDA topic modeling method. This algorithm has a built-in calculation of weight that finds the top words in each topic. Because of the repetitive nature of the job postings, future research would do well to examine the results arriving from other modeling tools. For example, comparing the LDA output with one from tmtoolkit would be illuminating. In addition, calculating the distinctiveness or saliency of the top words in each topic may lead us to better understand the importance of words and their contribution to topics.

As discussed above, the integration of English and Spanish-language postings presented problems as well. Future improvements might translate Spanish postings in place and deduplicate rows that were simply Spanish translations of English sections. However, our balance test above makes us comfortable that this exclusion did not seriously impact our results.

Finally, future analyses would benefit from restructuring the dataset in some ways. As described above, we calculated the weekly hours per job and would consider weighting future results by both the number of openings per job post and the hours per week for each opening.

These topic models can inform us about the types of H-2A jobs that get posted every year. A more refined topic model, linked to the violations dataset, may provide context clues that would flag employers for potential violations. The ability to streamline the investigation process would be a win for taxpayers, workers, and fair employers.

7 References

[1] R. Johnson. 2021. “Final Project: Social Impact Practicum.”

[2] “H2-A and H2-B Visa Programs: Increased Protections Needed for Foreign Workers.” United States Government Accountability Office: GAO Report to Congressional Committees, Washington, D.C. GAO-15-145, 2015.

[3] D. Costa and P. Martin, “Coronavirus and farmworkers: Farm employment, safety issues, and the H-2A guestworker program,” Economic Policy Institute, Washington, D.C. 2015.

[4] U.S. Citizenship and Immigration Services, “H-2A Temporary Agricultural Workers,” 2021.

[5] Center for Migrant Rights. “Recruitment Revealed: Fundamental Flaws in the H-2 Temporary Worker Program and Recommendations for Change,” Baltimore, MD. 2018.

[6] S. Li, “Topic Modeling and Latent Dirichlet Allocation (LDA) in Python,” Medium, 01-Jun-2018. [Online]. Available: https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24. [Accessed: 06-Jun-2021].

[7] D. Costa, P. Martin, and Z. Rutledge. “Federal labor standards enforcement in agriculture: data reveal the biggest violators and raise new questions about how to improve and target efforts to protect farm workers,” Washington, D.C. 2020.

[8] S. Bird, E. Loper, and E. Klein (2009), Natural Language Processing with Python. O’Reilly Media Inc.

[9] M.F. Porter, 1980, An algorithm for suffix stripping, Program, 14(3) pp 130137.