Using Knowledge Mining to Analyze H-1B Visa Sponsorship Trends in the U.S. Financial Sector from 2016 to 2024

Chakrayuddh Kimsovanna

EPPS 6323 – Knowledge Mining

Dr. Karl Ho

University of Texas at Dallas

May 12, 2025

Table of Contents

Introduction

Being one of the largest economies in the world, the United States has enjoyed decades of continuous growth across many sectors of industry, ranging from technology to commerce to manufacturing. As such, there has been a consistently increasing demand for high-skilled labor, and to help sustain this demand, the United States government relies on the H-1B visa program, which issues temporary nonimmigrant visas that bring in workers with specialized expertise in these areas from abroad. Where the home countries of these migrant workers might view this as brain drain, the United States sees this as a strategic advantage because it helps maintain a competitive edge for its economy. Currently, it has an annual statutory cap of 65,000 visas, plus an additional 20,000 visas for those who have recently graduate with a graduate degree from a United States institution (American Immigration Council 2025). It is argued that foreign workers actually contribute more to the United States workforce and economy than necessarily competing with native-born workers, as they would instead help filling in different types of job roles, boosting consumer demand, encouraging business growth and potentially introducing innovations that could help expand the American economy (American Immigration Council 2025).

While the majority of H-1B recipients work in fields like technology, engineering and healthcare, other sectors, including finance, also depend on skilled foreign professionals. Yet despite this, the finance industry is often overlooked in discussions about H-1B employment. When conducting background research through my university's database as well as Google Scholar, I found no studies focused specifically on H-1B sponsorship within the finance sector, although most of the existing literature tends to focus on STEM fields or on the program itself in general. To help address this gap in the research, this project uses knowledge mining methods to examine H-1B job trends within the finance industry to better understand how employer demand has shifted over time. The following sections will go over recent policy changes to the H-1B program, explain how the data was collected and analyzed, and then discuss the main findings, including some possible explanations for the trends.

Background

The H-1B visa program was first created in 1990 as part of the Immigration Act of that year to allow foreign workers to work in these specialty occupations, which are jobs that require a bachelor's degree and/or professional license in a specialized field, like engineering, science and academia (Batalova 2010). As mentioned earlier, the cap for these visas is still at 65,000 per year despite it briefly increasing to almost double that number in 1999. While the structure of the program has remained largely the same since, the political atmosphere surrounding it has changed a lot, especially over the last decade, starting with Trump's first presidential term. This is why this paper focuses on the period between 2016 and 2024, which includes Trump's and Biden's administrations, as both had very different policies and agendas, with almost opposing views on immigration, at least in terms of rhetoric. This time frame is also recent enough that the trends would still be relevant today, especially as ongoing debates over labor supply shortages and anti-immigration sentiment continue.

Within the first few months into President Trump's first term, he signed the Buy American and Hire American Executive Order, which tightened immigration enforcement and made visa programs like H-1B more restrictive (USCIS 2021). This order would ensure that H-1B visas would only go to the highest-paid or most-skilled applicants. In October 2020, the Trump administration made the process even harder by raising the minimum wage employers had to offer to these foreign workers and narrowing the definition of a specialty occupation so that less people would be qualified (Anderson 2020). This led to a surge in visa denials, but after President Biden took office in 2021, his administration reversed some of these changes, bringing denial rates down significantly compared the levels seen during Trump's term (Immigration Policy Tracking Project 2021).

However, in late 2023, the Biden administration signed a new order very similar to Trump's previous policy by restricting which jobs could count as specialty occupations (Anderson 2023). Nearing the end of his term, Biden made some final changes that allowed more flexibility with work start dates, officially extended F-1 student work authorization during the H-1B application process, and brought back deference to prior approvals; nonetheless, the rule still kept the same strict wording as before regarding specialty occupations (Shie et al. 2024). It is also important to note that the COVID-19 pandemic from 2020 to 2023 had a major impact on travel and with economic shutdowns and hiring freezes, it was very likely that H-1B visa numbers saw a major drop, too.

Methodology

Data Collection

This project mainly uses data from the United States Citizenship and Immigration Services (USCIS) and the Department of Labor (DOL), as they are both official and publicly available. Initially, data from job posting sites, such as LinkedIn, Handshake and Indeed, were considered for web-scraping, but due to their strict policy guidelines, I decided to just stick to these government data. The USCIS website contains the H-1B employer data from "fiscal year 2009 though fiscal year 2004 on employers who have submitted petitions to employ H-1B nonimmigrants workers" (USCIS 2024). It includes information about the approval and denial of H-1B visas, arranged into different sectors as defined by the NAICS code; here, finance would be in Group 52, lumping together with insurance, which I should keep in mind. Next is the performance data from the DOL, which contains all records for various types of visas; each file has the case status, Personally Identifiable Information (PII), as well as other related information, like wages, job titles, etc. (DOL2024). All these data come in CSV and XLSX format. The DOL files, in particular, are very large, and combined they total around 2.55 gigabytes in size, as each contains about hundreds of thousands of different cases.

Data Processing

Before building any model, I first had to clean the data to ensure consistency when analyzing multiple files. The USCIS data was relatively straightforward to work with, but the DOL data was much messier and required extensive cleaning; I wanted to filter by job titles, but they were all in unstructured and inconsistent formats, with random words and letters, as shown in Figure 1 below.

JOB_TITLE
Financial and Investment Analysts - KBGFJG72284-11
Model/Analysis/Validation Senior Analyst
Accountants and Auditors - Kbgfjg179876-5
Accountants and Auditors - Kbgfjg140954-6
Director, Sponsor Underwriting, Capital Markets
Treasury Analyst
Accountants and Auditors - KBGFJG102001-7
Associate III
FINANCIAL RISK SPECIALIST
Assistant Vice President
Technical Data Analyst
Accountants and Auditors - KBGFJG265299-1
Tax Senior
Accountants and Auditors - Kbgfjg227420-4
Supervisor, Audit & Attest
Audit Manager
Accountants and Auditors - Kbgfjg159935-6
Accountants and Auditors - Kbgfjg182835-4
Accountants and Auditors - Kbgfjg202575-4
Accountants and Auditors - KBGFJG101848-6
Accountants and Auditors - KBGFJG132073-4
Lead Trade Operations Analyst
Accountants and Auditors - Kbgfjg131909-4
Associate
Accountants and Auditors - Kbgfjg87152-10

Figure 1. Example of entries for the job titles from one of the DOL H-1B disclosure data.

I tried to filter out finance-related job titles directly within each Excel file, but found using RStudio to do so was much more efficient. First, I used the `readxl` package to import each XLSX file into R. Then, I used janitor to standardize the column names just in case there were any inconsistencies in formatting. After that, I used functions from the `dplyr` package to convert the texts to the format I wanted (e.g., changing everything under the variable `case_status` to lower case) and to apply a filter for cases that have either one of these values, "certified," "certified-withdrawn," "denied," or "withdrawn" while also getting rid of cases potentially without any job title.

```
finance_keywords <- c(
  "financ", "investor", "investing", "accounting", "accountant",
  "actuary", "treasury", "risk", "compliance", "quantitative",
  "portfolio", "equity", "asset", "securities", "hedge", "audit",
  "valuation", "credit", "bank", "underwriter", "tax"
)
```

Figure 2. Custom list of finance-related keywords used to filter H-1B job titles.

Admittedly, not all of these steps may have been strictly necessary, but I applied them out of caution, since I did not want to open all these large files to check. To filter out job titles specifically related to finance, I used the help of an AI tool, ChatGPT, to help me come up with all the possible names for these roles that might appear in the datasets. Figure 2 shown above lists all the keywords that I want to have; I later modified some of them, like "finance" and "financial," by combining into one "financ." These processing steps would allow me to reduce the size of each dataset significantly, so that when I create the models, the process was much faster. To save progress thus far, I converted each modified Excel file into an RDS file.

Embedding and Clustering Job Titles

Because the job titles varied widely in wording despite some referring to the exact same roles, it was difficult to group them accurately, so I used NLP techniques based on BERT, as discussed in class. To do so, I used Python within RStudio via the `reticulate` package. I imported the `sentence-transformers` library so that I could convert each job title into numerical vectors that can capture what each title means, making it easier to group semantically similar roles together. As a disclaimer, I relied on ChatGPT to help guide me and make sure I was doing these crucial steps properly. It recommended the use of the `all-MiniLM-L6-v2` model, which "is a compact but powerful model [that] takes a sentence or a short paragraph and maps it to a 384-dimensional vector, making it easy to use for tasks such as [semantic search, clustering and sentence similarity]" (Tiwari 2024). Before I generated the embeddings, I further cleaned the job titles by removing unnecessary words like "SR," "junior," "president," "director," "associate" and "CEO" because in this analysis, I was not interested in the rank of the position, only the role itself. I also removed any numerical value and unnecessary punctuation in those job titles. To make sure I achieved the same consistency for each dataset, I combined all the cleaned DOL files from 2016 to 2024 into one list, which I then put into the embedding model to convert each job title into a 384-dimensional vector, as titles with similar meanings would be placed near each other in that vector space. After I created these embeddings, I saved the datasets for each year separately, in case I needed to go back and analyze specific time periods later. I also combined all the unique titles across the years into a single list and saved both the text (`dol_h1b_combined_titles.rds`) and their corresponding embeddings (`dol_h1b_embeddings.rds`). The former contains the job titles themselves, while the latter stores the vectors as described earlier.

To create clusters to group these vectors together, I first had to determine the optimal number of clusters, *k*, using the silhouette score method. Using the `sklearn.metrics` library, I calculated the scores for values of *k* from 5 to 25, running them on a random sample of 1,000 embeddings instead of the full dataset, since that would have been too intensive for my computer. As shown in Figure 3 below, the highest silhouette score is at $k = 18$. Although the score was quite low at just above 0.07 (far from being close to the ideal 1.0), it was still the best result. Initially, the highest score I got was $k = 8$. However, because I made a slight modification to the 2020-2024 datasets (specifically correcting the filtering to include "certified - withdraw" cases as I had previously only included "certified-withdraw"), the silhouette analysis now showed this instead. However, by this point, I just chose to keep it at 8 in order to maintain consistency with the outputs I had already generated and to keep the analysis more interpretable. Additionally, k being equal to 8 still produced the highest score within the lower range of cluster values, before the sharp increase beyond k =15.

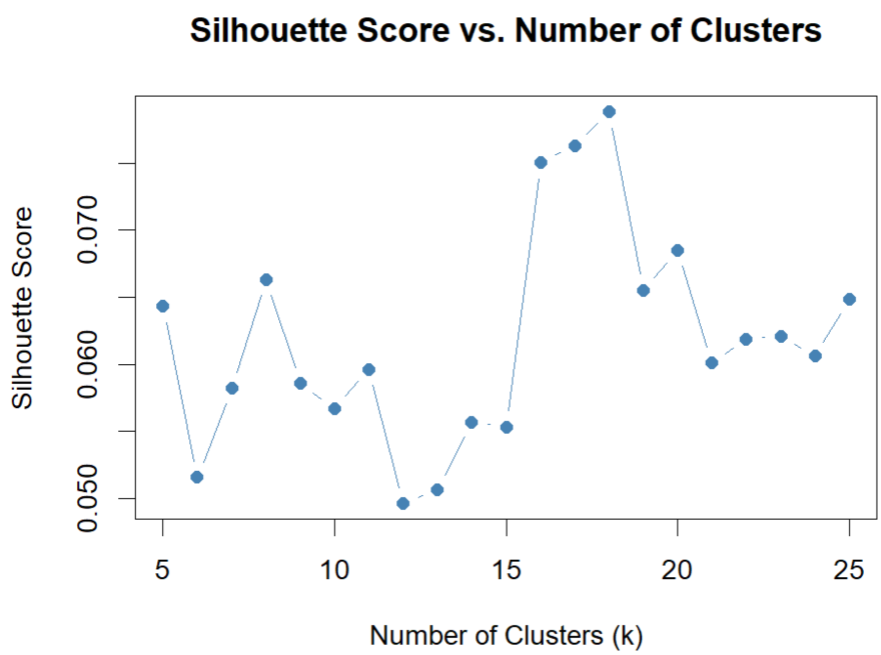

Figure 3. Plot showing the silhouette score for optimal number of clusters, k.

Afterwards, I ran KMeans clustering using `sklearn.cluster` with $k = 8$ on the job title embeddings. This process would assign each job title to one of the eight clusters based on their semantic similarity, with each cluster temporarily represented by a number from 0 to 7. I then updated all yearly datasets to have these cluster values as well. After that, I combined all the datasets into one final CSV file (`dol_all_tc_with_cluster_status.csv`) for analysis. I also would like to mention that I used another AI tool, Copilot, to not necessarily help me write the codes, but to check the syntax errors that I had frequently gotten into when coding these parts. Nevertheless, to help interpret what each cluster represents, I used the `dplyr` package to count and extract the top 50 most frequent job titles within each cluster. While this showed the most common roles, it did not fully reveal the overall wording patterns across titles.

Visualization of Job Clusters

One method that could visualize each cluster is a word cloud, which displays the most frequently occurring terms, as the more frequent those words appear, the larger they are shown. Here, I used the `wordcloud` package in R to generate a word cloud for each cluster based on its respective set of top job titles. After noticing some strange and irrelevant terms like "and," "i" and "ii" when generating the word clouds, I added additional filters to remove these as well. Instead of having eight separate word clouds, I decided to combine them into a big panel.

Results and Discussion

This section goes over each of the eight clusters, providing additional details and possible explanations for why certain trends appeared over time. As one can see in Figure 4 below, each cluster contains a few standout terms, and fortunately, these dominant words are mostly distinct across clusters, except for some terms like "finance" and "financial" which could be due to them being part of the names for those jobs—for example, "financial banking specialist" and "financial data analyst."

Figure 4. Panel of word clouds showing the most common job title keywords for each cluster.

This panel of word clouds made it easier to infer what each group represents, which I have summarized in the table below:

Cluster	Finance Job Category
0	Banking and credit services
1	Financial systems and support services
2	Risk management
3	Tax, audit and accounting
4	Financial and business analytics
5	Quantitative and research analytics

6	Portfolio and asset management
7	Compliance and regulation

Table 1. Labels for each cluster based on word cloud results.

Ignoring the words “financial” and “finance,” Cluster 0 was labeled as “banking and credit services” due to the prominence of terms like “banking” and “credit.” The same happened to Cluster 1 with the presence of the term “finance;” however, I labeled it as “financial systems and support services” because the cluster includes terms like “consultant,” “services,” “engineer” and “specialist,” which all suggest roles related to operational support and technical assistance. Cluster 2 centers around “risk” so it would be “risk management.” Cluster 3 would be “tax, audit and accounting,” which makes sense because these roles often overlap as they involve financial reporting and preparation of financial statements. Cluster 4 and 5 are very similar to one another as they both involve analytics; the only difference is that the former contains words like “business” while the latter focuses more on “quantitative” and “research,” hence the labels are now “financial and business analytics” and “quantitative and research analytics,” respectively. For Cluster 6, words like “portfolio” and “asset” stand out, so it would be categorized as “portfolio and asset management.” Additionally, I labeled Cluster 7 as “compliance and regulation”, since “compliance” was by far the most dominant term and I also noticed “regulatory,” which appeared smaller but still relevant. This cluster clearly points to roles focused on overseeing rules, policies and regulatory requirements.

The following summary table shows the total number of jobs recorded for each cluster from 2016 to 2024, which includes all those that were certified, withdrawn or denied (Table 2). The most dominant group throughout this period is Cluster 3 (tax, audit and accounting), followed by Cluster 4 (financial and business analytics) and Cluster 2 (risk management). In contrast, the clusters with the fewest job titles overall are Cluster 6 (portfolio and asset management) and Cluster 7 (compliance and regulation), both of which remained consistently low across all years. The other clusters that are in mid-range include Cluster 0 (banking and credit services), Cluster 1 (financial systems and support services) and Cluster 5 (quantitative and research analytics).

	Years								
Cluster	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**	**2024**
0	1578	1675	1895	1863	1748	2931	2188	2338	2175
1	1336	1349	1517	1465	1209	1787	1279	1710	1762
2	1476	1553	2185	2212	2001	3141	2117	1869	1746
3	8468	8907	9372	9504	8189	11411	7973	10110	8051
4	3680	3737	4287	4097	3262	5569	3707	4222	3843
5	1504	1741	2041	2229	2133	3042	2092	2291	2157
6	511	506	650	616	567	1008	678	697	767
7	579	656	693	729	721	1034	784	795	774

Table 2. Total number of jobs by cluster and year, including all case statuses (2016-2024)

To visualize the trends of each cluster, including the breakdown by case status, I used the `ggplot2` package to create bar charts for each of them, which displays the yearly counts of certified, certified-withdrawn (case where the petition was approved but later withdrawn), denied and withdrawn cases from 2016 to 2024.

As one can see in the bar graph below, showing the first cluster (banking and credit services), the total number of cases gradually increased from 2016 to 2019, followed by a slight decline in 2020 and a sharp spike in 2021 (Figure 5). This was about 86 percent growth from

2016 to 2021. However, after the COVID-19 pandemic (2021 to 2024), the total number of cases decreased by about 26 percent, though it still remained higher than the pre-pandemic levels. The proportions of other case types remained surprisingly constant regardless of the number of certified throughout the entire period, except in 2020 when cases that were certified-withdrawn dropped to less than half of their usual levels.

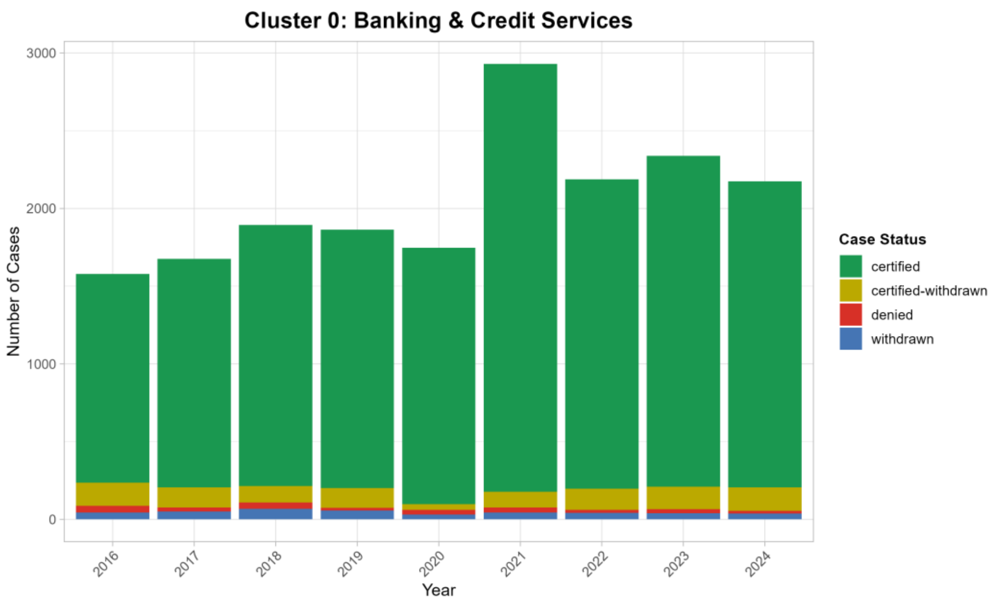

Figure 5. Graph showing number of cases by year and status for Cluster 0 (banking and credit services).

For Cluster 1 (financial systems and support services), there are similar patterns to the previous one in terms of the total number of cases, where there was a 48 percent increase from 2016 to 2021, before it dropped by about 28 percent the year after (Figure 6). However, from then onwards, the number began to rise again. In addition, there were fewer denials and withdrawals since 2020 compared to the earlier years.

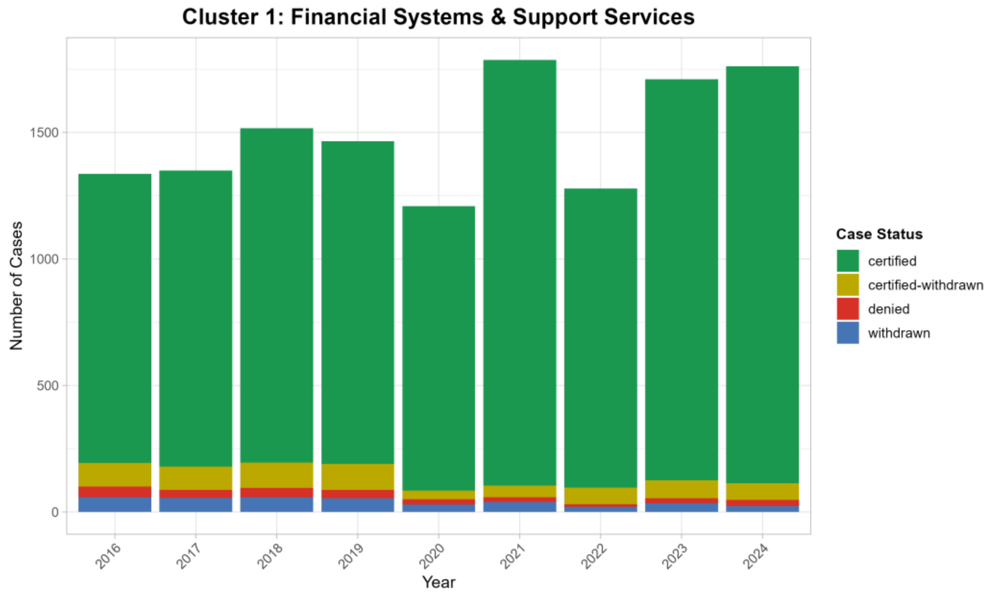

Figure 6. Graph showing number of cases by year and status for Cluster 1 (financial systems and support services).

The overall distribution of Cluster 2 (risk management) is more like a normal distribution, with a steady rise from 2016 to 2021 at a growth of 113 percent, and a gradual decline afterward, falling by roughly 44 percent (Figure 7). The proportion of denials and withdrawals were much lower than those from the previous two clusters.

Figure 7. Graph showing number of cases by year and status for Cluster 2 (risk management).

Being the largest group, Cluster 3 (tax, audit and accounting) consistently had the highest number of total cases across the observed period (Figure 8). In 2016, there were about 8,468 total cases petitioned, and this number would increase slightly until 2020 when it dropped to 8,189. In 2021, it grew by almost 40 percent, before it dropped again in 2022 to a level similar to 2020. It increased again in 2023, before dropping in 2024 to that same level as in 2020 and 2022.

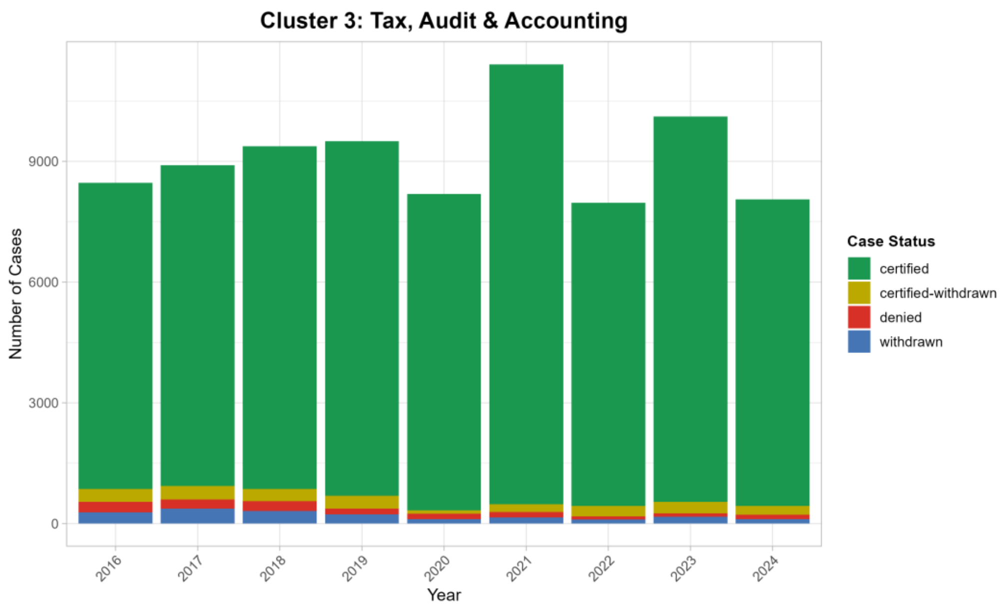

Figure 8. Graph showing number of cases by year and status for Cluster 3 (tax, audit and accounting).

In Figure 9 shown below, we can see that the number of cases for Cluster 4 (financial and business analytics) was mostly stable, hovering between 3,600 and 4,300 cases throughout the years, except in 2020, when it was at just 3,262, as well as 2021 where the number went up to 5,569.

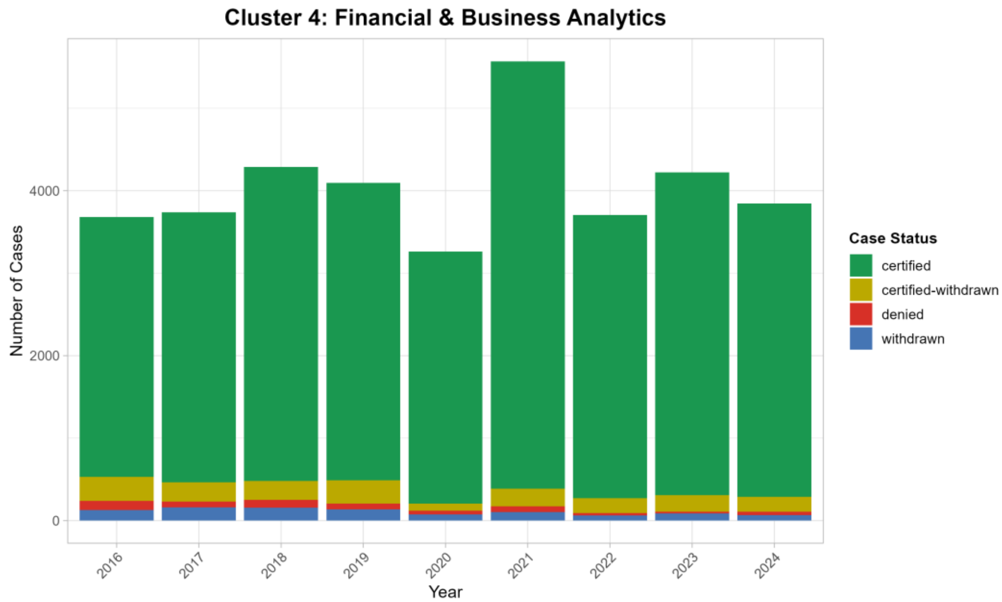

Figure 9. Graph showing number of cases by year and status for Cluster 4 (financial and business analytics).

Like Cluster 4, Cluster 5 (quantitative and research analytics) followed a similar trend, with total cases mostly increasing until 2021, despite a slight dip in 2020, which marked an overall increase of 102 percent (Figure 10). In 2022, the number dropped back down by 31 percent to 2,092 at about the same level as in 2020, and remained relatively stable through 2024.

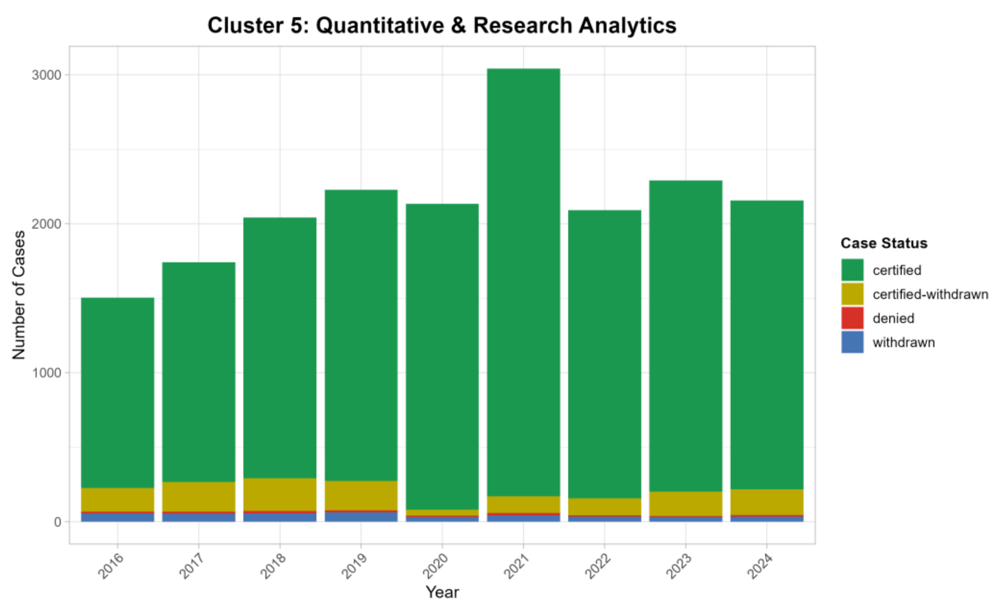

Figure 10. Graph showing number of cases by year and status for Cluster 5 (quantitative and research analytics).

Cluster 6 (portfolio and asset management) had the lowest number of cases, starting with just 511 in 2016 and remaining around that level until 2020 (Figure 11). In 2021, again, the number jumped to 1,008, which was almost double the previous year, but then dropped back to 678 in 2022. After that, it slowly increased again in 2023 and 2024, reaching 767. Despite the smaller totals, the general pattern of a peak in certified cases in 2021 followed by gradual recovery along with the distribution across case statuses (i.e., withdrawals and denials remaining mostly constant throughout, with just 2020 showing the lowest number of certified-withdrawals), was consistent with all the other clusters thus far.

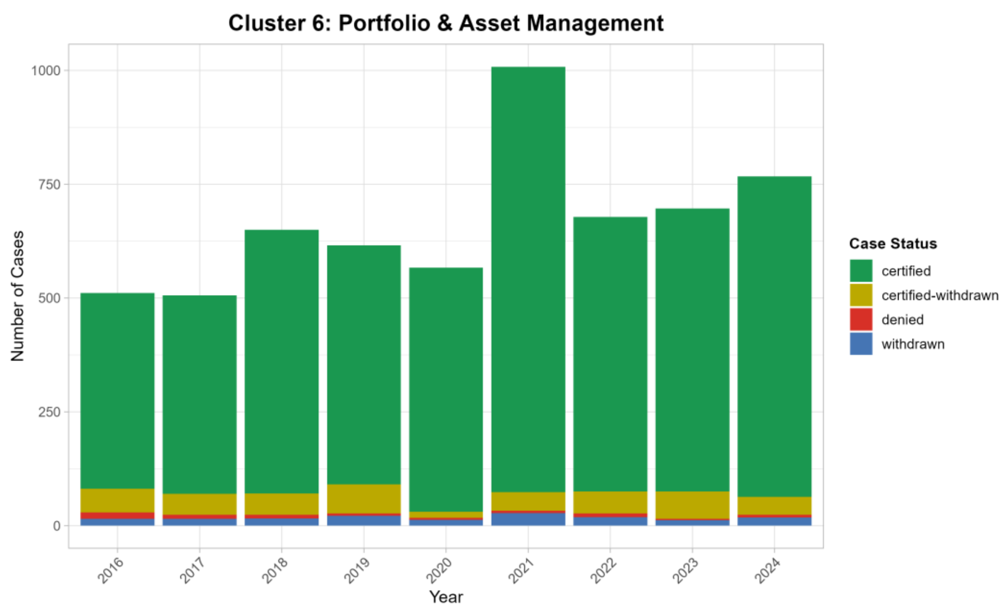

Figure 11. Graph showing number of cases by year and status for Cluster 6 (portfolio and asset management).

Lastly, Cluster 7 (compliance and regulation) being the second smallest group, experienced a modest upward trend of approximately 34 percent judging from the numbers in 2024 and 2016 (Figure 12). Again, the number rose drastically in 2021, then leveled off in the following years, remaining slightly above the 2020 figure.

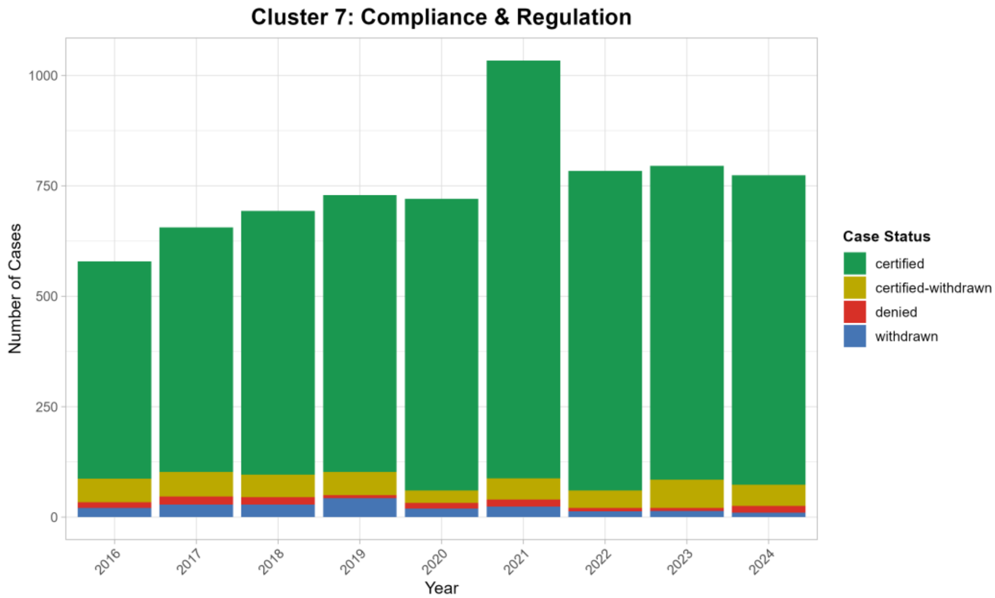

Figure 12. Graph showing number of cases by year and status for Cluster 7 (compliance and regulation).

The explanation for why 2021 stands out as an anomaly, as all of these job categories saw a sudden spike in cases, very likely comes down to Biden's administration loosening some of the restrictive immigration policies implemented under Trump, as described earlier. The subsequent decline in case numbers can be attributed to the COVID-19 pandemic, as there were travel restrictions, as well as uncertainty and instability in the economy, which led to many companies to freeze any new hiring. Even though the pandemic had mostly subsided by 2024, its lingering effects on labor demand likely continue to influence visa sponsorship levels. This was further compounded by the Biden administration later in its term reintroducing several of Trump's restrictive measures. This broader trend is further supported by the USCIS data, as depicted in Figure 13 below, which shows a drastic increase in continuing approvals between 2020 and 2022, meaning that H-1B workers were receiving extensions and renewals at a higher rate, as reflected by the number of certified petitions in the previous graphs. Surprisingly, the number of initial approvals did not increase much during this specific time period. Instead, it remained stable and only began to show a clear sign of increase from 2023 to 2024 (while continuing approvals declined). This trend may suggest that employers were more likely to focus on retaining existing H-1B workers during the more uncertain years (COVID-19 era) and only resumed new sponsorships once economic conditions began to improve. As for the denials—both initial and continuing—they remained very low and relatively constant throughout the entire period.

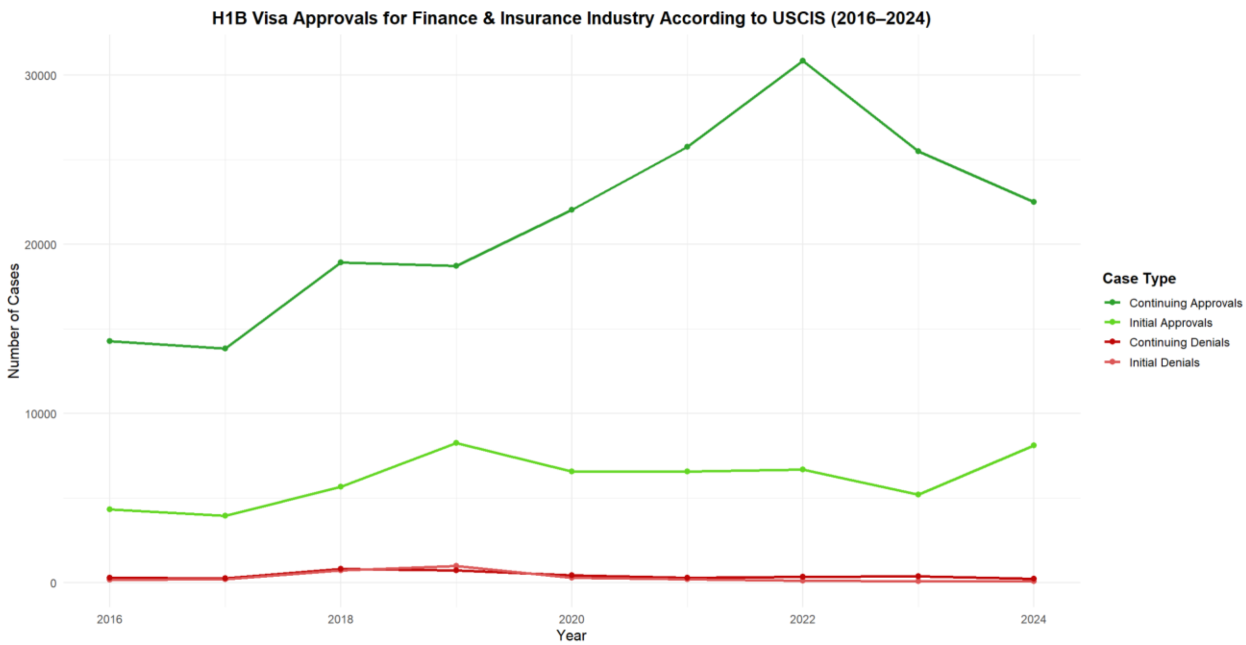

Figure 13. Graph showing the number of approvals and denials for H-1B visas in the finance and insurance industry (NAICS 52).

Challenges and Limitations

Over the course of this project, I encountered many difficulties, particularly during the stages of data cleaning and job title clustering. Due to the massive size of the datasets, processing and transforming the data required significant memory resources from my computers. My laptop frequently crashed during these tasks, and although I was able to rely on my desktop PC as a backup for the more intensive work, even that system experienced multiple freezes, occasionally requiring hard resets. These technical limitations did slow my progress but I managed to circumvent these issues, such as by using a random sample of 1,000 job title embeddings during the silhouette scoring process to estimate the optimal number of clusters, rather than the full set. Even the data itself, particularly regarding the job titles, may not be entirely accurate, as much of it was unstructured. It is also possible that some job titles that should have been excluded, as well as those that should have been included, may have been inadvertently misclassified or omitted due to the filtering process potentially not being robust enough.

Moreover, I initially planned to analyze the wages and salaries for each of the jobs, but this proved very difficult because they were reported in varying units (e.g., hourly, weekly, monthly and annually). I tried to standardize all entries to an annual format, but the process often resulted in some jobs with implausibly large values (sometimes in the millions) possibly due to calculation errors or the codes not handling certain cases correctly. As a result, I decided to abandon the wage analysis altogether and instead focused solely on approval rates and overall job demand for each of these finance job categories.

Conclusion

This study used knowledge mining techniques to examine how demand for H-1B jobs in the United States' financial sector has changed over the past decade in response to changing policies across different administrations and other external factors like the COVID-19 pandemic. By applying semantic clustering with BERT, it was possible to identify distinct categories of finance roles, such as accounting, risk management, and quantitative and research analytics, and analyze how demand for each has changed relative to the others over this period. Future research can build upon this by incorporating additional years of data or even expanding the analysis to include other sectors beyond finance. Including more reliable wage information and perhaps geographical details showing how different states compare, could provide a more comprehensive understanding of how job demand in this sector varies as well. Finally, the filtering and clustering processes could be further improved by using more robust models and functions to allow for more accurate grouping of unstructured data and more reliable results overall.

References

American Immigration Council. 2025. "The H-1B Visa Program: A Primer on the Program and Its Impact on the U.S. Economy." *American Immigration Council*. https://www.americanimmigrationcouncil.org/research/h1b-visa-program-fact-sheet.

Anderson, Stuart. 2020. "Trump Administration Issues Two New Rules to Restrict H-1B Visas." *Forbes*, October 7. https://www.forbes.com/sites/stuartanderson/2020/10/07/trump-administration-issues-two-new-rules-to-restrict-h-1b-visas/.

Anderson, Stuart. 2023. "Biden Immigration Rule Copies Some Trump Plans to Restrict H-1B Visas." *Forbes*, October 23. https://www.forbes.com/sites/stuartanderson/2023/10/23/biden-immigration-rule-copies-some-trump-plans-to-restrict-h-1b-visas/.

Batalova, Jeanne. 2010. "The H-1B Temporary Skilled Worker Program." *Migration Policy Institute*. https://www.migrationpolicy.org/article/h-1b-temporary-skilled-worker-program.

Department of Labor (DOL). 2024. "Foreign Labor Performance Data." *U.S. Department of Labor*. https://www.dol.gov/agencies/eta/foreign-labor/performance.

Shie, Grace, Maximillian L. Del Rey, and Morgan Bailey. 2024. "H-1B Modernization: Biden's Final Move, Trump's First Challenge." *Mayer Brown*, December. https://www.mayerbrown.com/en/insights/publications/2024/12/h1b-modernization-bidens-final-move-trump-first-challenge.

Tiwari, Rahul. 2024. "Unlocking the Power of Sentence Embeddings with All-MiniLM-L6-v2." *Medium*. https://medium.com/@rahultiwari065/unlocking-the-power-of-sentence-embeddings-with-all-minilm-l6-v2-7d6589a5f0aa.

U.S. Citizenship and Immigration Services (USCIS). 2021. "Buy American and Hire American: Putting American Workers First." *USCIS*. https://www.uscis.gov/archive/buy-american-and-hire-american-putting-american-workers-first.

U.S. Citizenship and Immigration Services (USCIS). 2024. "H-1B Employer Data Hub." *USCIS*. https://www.uscis.gov/tools/reports-and-studies/h-1b-employer-data-hub.

U.S. Immigration Policy Tracker. 2021. "Report That USCIS Has Increased Rates of RFEs and Denials of H-1B and L-1 Petitions." *Immigration Policy Tracking Project*. https://immpolicytracking.org/policies/report-that-uscis-has-increased-rates-of-rfes-and-denials-of-h-1b-and-l-1-petitions.