Trends in T-Visa Applications

Yizhou Mao, Zixuan Jiang, Heriberto Varela

Contents

1 Introduction

This project will look at T-Visas, a visa program for victims of human trafficking. There are 5,000 T-Visas available each year, however, never have more than 1,000 been granted in a year. Additionally, there has been an apparent increase in denials in T-Visa applications (there are 5,000 available per year). There is a hypothesis that the increase in denials is related to procedural barriers that immigration is constructing to make it more difficult for applications to find success. To evaluate this hypothesis, we have analyzed past decided applications and looked for trends in their decisions. Specifically, our team looks at administrative appeal decisions (AAO) of motions to reopen and reconsider. The analysis performed primarily consisted of aggregating the available data by year, and looking for keywords in the description and the explanation of the decisions.

2 Data Collection & Analysis

At the beginning of the project, we intended to obtain the T-Visa applications from the United States Citizenship and Immigration Services (USCIS) directly. Given some limitations of the search engine, we were only able to obtain a subset of the data needed, so we ended up completing the project with data provided directly by the client. We will explain the processing of the data for both stages, first the scraping the data from the USCIS website, and second the processing of the data provided by the client.

2.1 USCIS Website

2.1.1 Data source from USCIS website (AAO Non-Precedent Decisions)

Firstly, we need to download all the required documents from the website by sending different HTTP requests to the USCIS website and download them into a local folder.

The applications are in PDF format, so first we needed to download all the pdf files from the USCIS website. We noticed the URL had a certain pattern, so we used this pattern and concatenated the URL with the corresponding page number to retrieve each application PDF online. There are two Python libraries available that were relevant to our work, PyPDF2 and pdfminer. PyPDF2 was our default choice, since pdfminer could only extract data without formatting.

For preprocessing of the data in each application PDF, we used the PdfFileReader method from PyPDF2. This method helped us extract information such as numPages (number of pdf pages), and then we extracted the data from the PDF as text from each page iteratively, appending it into a separate array for analysis later.

2.1.2 Data Preparation

At this point we had the texts from all the available application PDFs, so we started to extract the information we were going to use for analysis. The following list summarizes the general extraction process:

1. Search for texts that had an ID number by using the search method included in the library.
2. From such texts, we extracted data such as decision, status, order, is_family, and description, values that are explained in the documentation of the project. This was done using keywords that identify specific sections of the application, which were provided by the BU Law team.
3. We stored the values for file_name, url, and date. These values were not immediately available from the texts, so these were obtained from the URL of each file.
4. Finally, we organized the data, sorted it by date, and converted the dataset into a CSV file.

2.2 LEXIS Database

2.2.1 Data source from client (LEXIS database)

As mentioned before, given the limitations of the USCIS website, the client provided us the full dataset needed for the project. The applications were still in PDF format, however the sections and the formatting of the documents were different, so we had to adjust our preprocessing methods. Generally, unlike the files from the USCIS website, these files did not have a fixed format. Thus, most of the extraction of data had to be done using keywords. By extracting the important information from these parsed texts, we employed the regular expressions from re package. This package allows us to extract the data from text based on the certain pattern. Then, we can insert all these data into our table and save them as csv file in the local folder.

2.2.2 Data Preparation

We parsed each file using PyPDF2 and the following lists explain the extraction of the data: Define all the variables that will store the extracted values. Iterate through the application PDFs in order and process them separately. Remove all the hidden files such as .DS_Store.

Extract the name of the PDF and store them in variable file_name. Extract regular information like id, date, decision, status, order, is_family, description, citation. Detailed steps are described at the end of the report [1].

2.3 Data Analysis

As mentioned above, our analysis consisted in analyzing keywords in the description and the explanation of the applications. We specifically took two approaches: we measured the frequency of a list of keywords provided by the BU Law team, and we extracted the “top” words in the description and explanation of the application. The specifics of both approaches are highlighted below:

2.3.1 First Approach - Keyword frequency

The BU Law team provided us with the following list of keywords and phrases: "victim of a severe form of trafficking in persons", "physically present", "complied with reasonable requests for assistance", "extreme hardship", "inadmissible" and "inadmissibility".

The measurement of the frequency of these keywords consisted in checking for their prevalence in the fields of our data containing the description of the application (‘description’), and the ANALYSIS/OPINION sections (‘contents’). We registered how many times each keyword was contained within these fields for each type of decision in the applications.

2.3.2 Second Approach - Top words

For each decision, we believe there must be keywords that can characterize it and its applicant, so we performed the top words analysis on the description and the explanation of the applications of each decision.

For preprocessing the description and the explanation, we lemmatized each word and removed the words that are stop words or are not related to decision, such as applicant, application, and document. We also chose to include only words that are nouns because including verbs and adjectives did not produce meaningful results. We then utilized the CountVectorizer from Sklearn to count the appearance of each word and sorted them from highest to lowest. By visualizing the top words of each decision with graphs, we found some characteristics that might be indicative of certain decisions.

3 Results & Key Insights

3.1 How many cases were decided?

As mentioned above, we handle the motions to reopen and reconsider, which are the applications for motions on the decisions done by the Administrative Appeals Office. From the processed data, We were able to obtain a total of 32 decided motion cases by filtering them out using Pandas. We then counted the number of cases for each decision and utilized matplotlib.pyplot to generate graphs for each decision. Figure 1 shows the total amount of motion cases per year. Figure 2 shows the number of motion cases by decision.

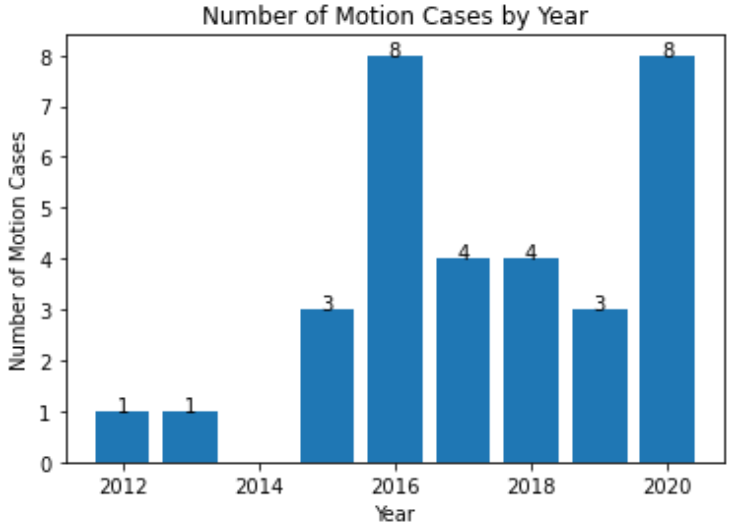

Figure 1: Number of Motion Cases by Year

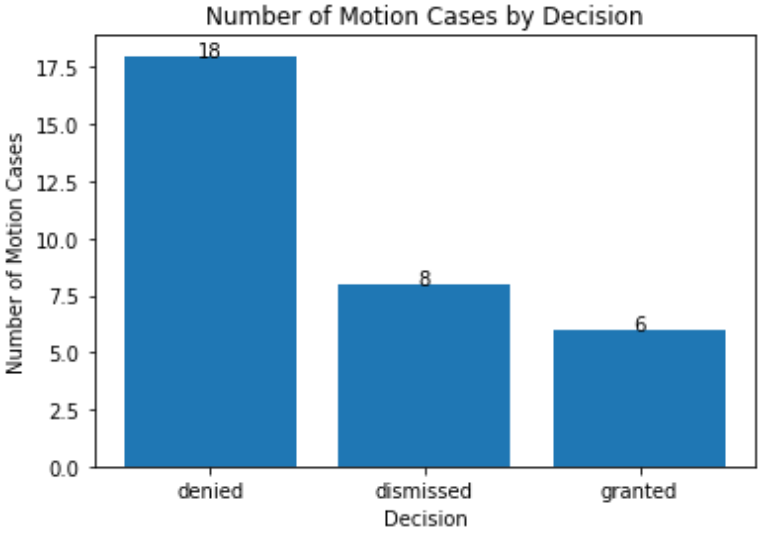

Figure 2: Number of Motion Cases by Decision

Figure 1 shows that the earliest Motion case on the dataset is from 2012, and it was the only one from that year. Similarly, 2013 also shows only 1 Motion case. 2014 did not present any Motion cases but beginning in 2015 we started seeing more cases. 2015 presented 3 Motion cases, 2016 presented 8 Motion cases, both 2017 and 2018 presented 4 Motion cases, 2019 presented 3 Motion cases, and finally 2020 presented 8 Motion cases.

Figure 2 shows that the most prevalent decision for the Motion cases was denial, with 18 denied Motion cases. 8 Motion cases were dismissed, and only 6 were granted.

3.2 How many cases were dismissed?

Out of the 32 decided motion cases, 8 were dismissed. Shown in Figure 3 are the dismissed motions by year.

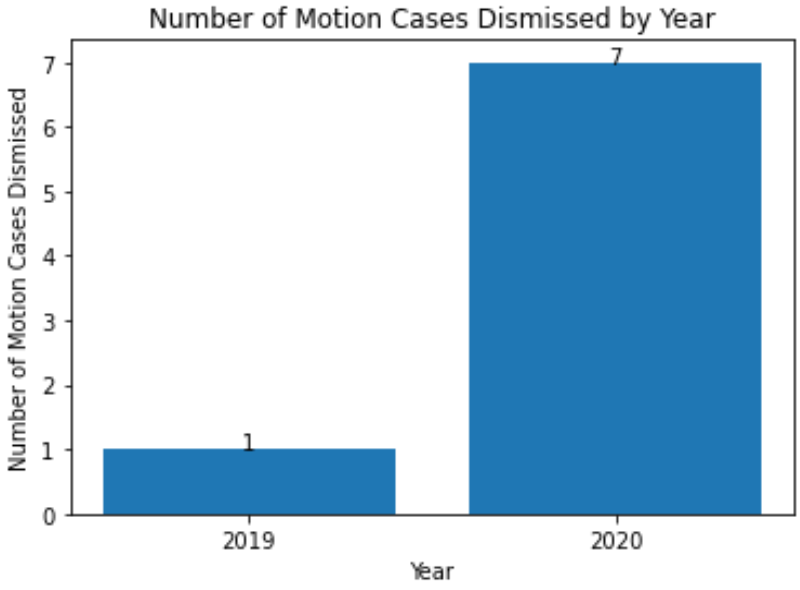

Figure 3: Number of Motion Cases Dismissed by Year

Figure 3 shows that 2019 presented 1 dismissed Motion case, and 2020 presented 7 dismissed Motion cases.

3.3 How many cases were granted or denied?

Out of the 32 decided motion cases, 6 were granted and 18 were denied. Shown in Figure 4 are the granted motion cases by year. As we can see, only 1 or 2 motions were granted per year in the presented years. In Figure 5, we can see the denied motion cases by year.

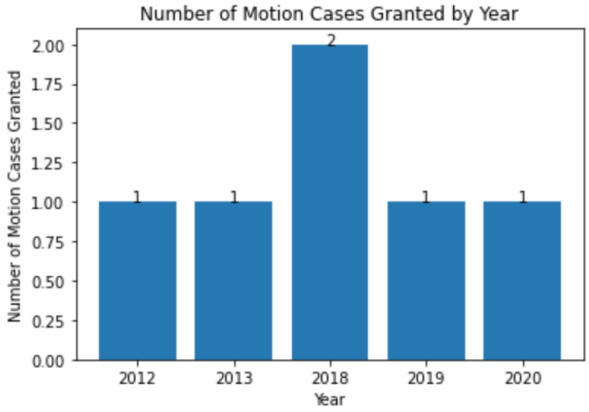

Figure 4: Number of Motion Cases Granted by Year

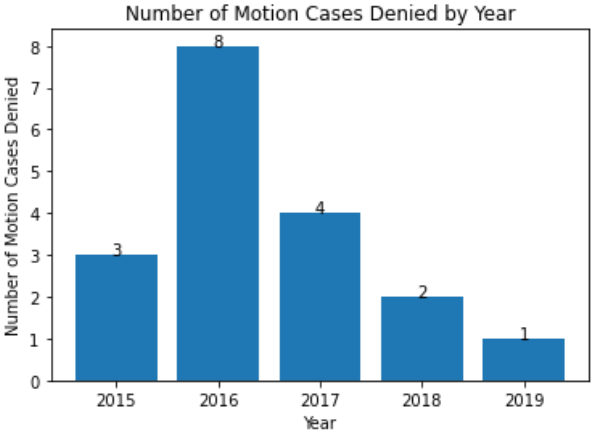

Figure 5: Number of Motion Cases Denied by Year

Figure 4 shows that 2012, 2013, 2019 and 2020 presented 1 granted Motion case, and that 2018 was the only year that presented 2 granted Motion cases. Figure 5 shows that 2015 presented 3 denied Motion cases, 2016 presented 8 Motion cases, the most out of all decided cases, 2017 presented 4 denied Motion cases, 2018 presented 2 denied Motion cases, and 2019 presented 1 denied Motion case.

3.4 On what grounds where cases decided?

As mentioned before, to examine the grounds of the decisions, we implemented two different approaches. First, we looked at the frequency of the keywords provided by the supporting BU Law team in the application's description.

On Figure 6, we can see the occurrence of the keywords in granted Motion cases. On Figure 7, we can see the occurrence in denied Motion cases. On Figure 8, we can see the occurrence in dismissed Motion cases, the full keywords analyzed were: "victim of a severe form of trafficking in persons", "physically present", "complied with reasonable requests for assistance", "extreme hardship", "inadmissible" and "inadmissibility". The keywords omitted in the figures had an occurrence of 0.

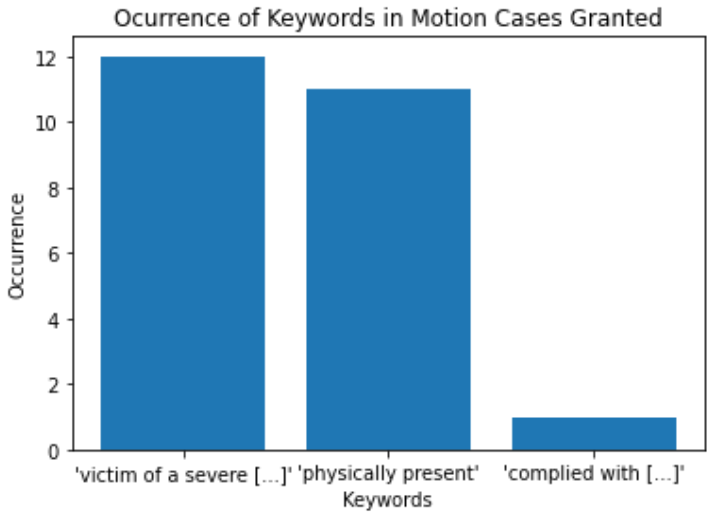

Figure 6: Keyword Frequency in Motion Cases Granted

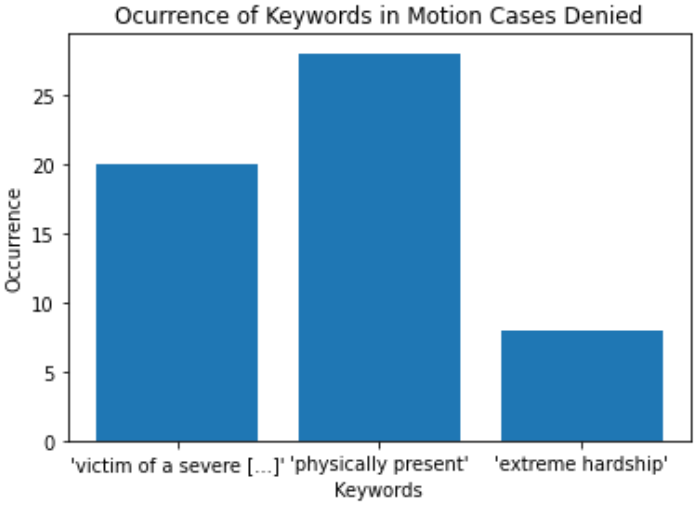

Figure 7: Keyword Frequency in Motion Cases Denied

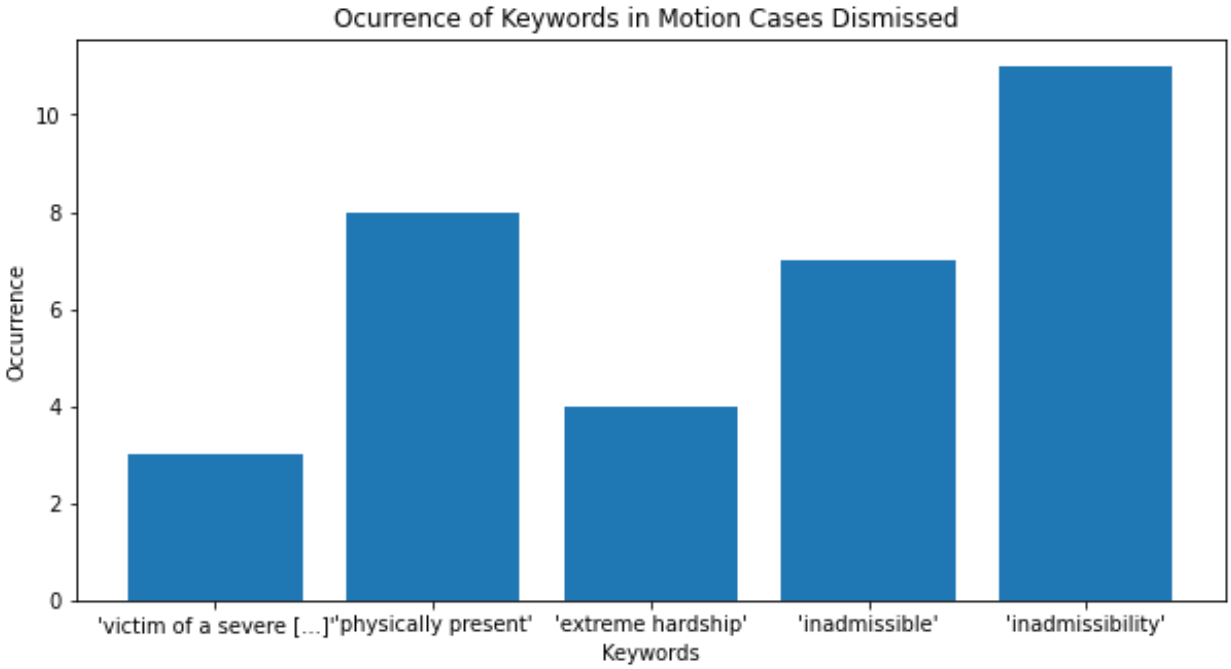

Figure 8: Keyword Frequency in Motion Cases Dismissed

Figure 6 shows that for the granted Motion cases, the most frequent keyword out of the provided list was "victim of a severe form of trafficking in persons", with 12 matches in the applications. In descending order of frequency, the following keywords were "physically present" with 11 matches found, and "complied with reasonable requests for assistance" with 1 match found.

Figure 7 shows that denied Motion cases, the most frequent keyword out of the provided list was "physically present", with 28 matches in the applications. In descending order of frequency, the following keywords were “victim of a severe form of trafficking in persons” with 20 matches found, and "extreme hardship" with 8 matches found.

Figure 8 shows that for the dismissed Motion cases, the most frequent keyword out of the provided list was "inadmissibility", with 11 matches in the applications. In descending order of frequency, the following keywords were "physically present" with 11 matches found, "physically present" with 8 matches found, “inadmissible” with 7 matches found, “extreme hardship” with 4 matches found, and "victim of a severe form of trafficking in persons" with 3 matches found.

As for the second approach, we looked at the top words with the most frequency in the application’s Analysis and Opinion sections. In Figure 9, we can see the frequency of the top words on denied Motion cases. In Figure 10, the frequency on granted Motion cases, and in Figure 11 the frequency on dismissed Motion cases.

Figure 9 shows the top 15 words in motion cases that were denied. There are some words that might indicate the denial of the cases, such as violence, drug, and employment. Thus, if an applicant has characteristics related to violence, drug, and employment, his/her motion to reopen and reconsider is more likely to be denied.

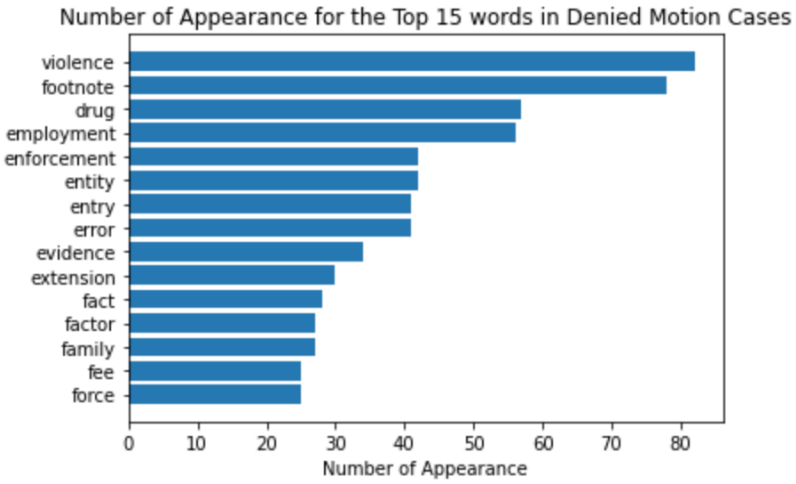

Figure 9: Frequency Graph of Top Words in Denied Motion Cases

Figure 10 shows the top 5 words in motion cases that were granted. There are some words that might indicate the grant of the cases, such as victim, trafficking, and counsel. Thus, if an applicant has characteristics related to victim and trafficking and is represented by counsel, his/her motion to reopen and reconsider is more likely to be granted.

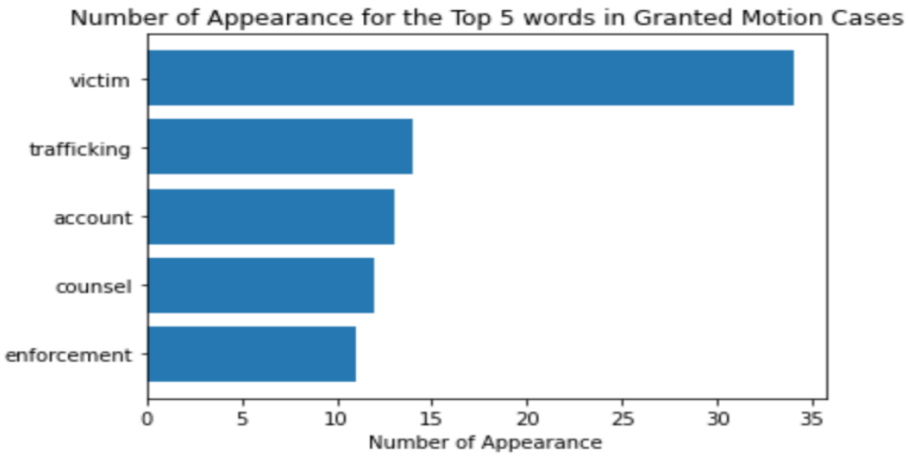

Figure 10: Frequency Graph of Top Words in Granted Motion Cases

Figure 11 shows the top 15 words in motion cases that were dismissed. There are some words that might indicate the dismissal of the cases, such as record and waiver. Thus, if an applicant has some kinds of records and is considered as a waiver, his/her motion to reopen and reconsider is more likely to be dismissed.

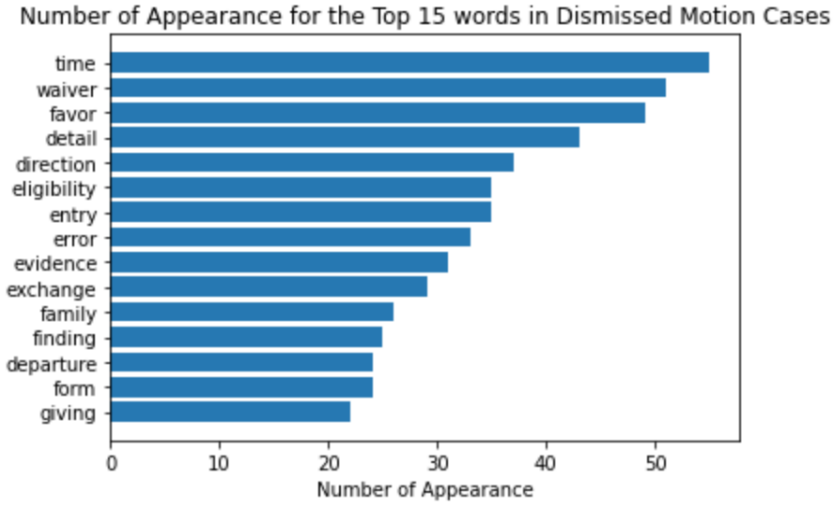

Figure 11: Frequency Graph of Top Words in Dismissed Motion Cases

3.5 How has this changed over time?

In Figure 1, we can see that 2016 and 2020 were the years when most motions were decided, with 8 applications each. In Figure 4, we can see that the amount of granted motion cases did not vary across the years with only 1 or 2 cases each year, and in Figure 5 we saw that 2016 was the year when most motion cases were denied. Another interesting insight found in Figure 3 is that 2020 was the year with the most dismissed motion cases.

3.6 Are there any patterns?

On the prevalence of the keywords provided by the BU Law team, we found interesting that the keywords "victim of a severe form of trafficking in persons" and "physically present" were prevalent in all Motion decisions. The difference in the prevalence of keywords presented in Figure 6, 7 and 8 could be an indicator of the importance of these terms in their respective Motion decisions. On the frequency of words, if we disregard words that are of standard use in the applications such as ‘trafficking’ and ‘victim’, we can see on Figure 9, Figure 10 and Figure 11 what could also be important terms related to the Motion decisions.

Additionally, we were able to extract whether an application had a “Counsel” section, indicating that an applicant was assisted with their representation. On Figure 12, we can see the number of applications that included a Counsel section.

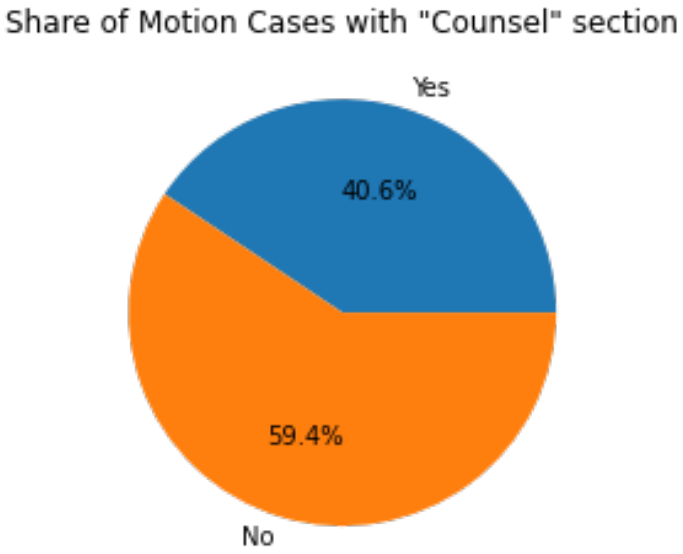

Figure 12: Share of Motion Cases with Counsel section

Figure 12 shows us that the majority of the Motion cases did not include a Counsel section, with 19 Motion cases (59%) not including it, and 13 Motion cases (40%) including it.

4 Limitations and Future Work

The most difficult part of this project was the data processing. After receiving the final dataset from the client, we had to adapt the processing to handle PDFs with different formats. Another limitation we encountered was the size of our dataset. We were unable to apply more comprehensive analysis tools given that we only had 32 valid applications.

We hope our findings help the BU Law team on the evaluation of the project hypothesis, and have highlighted some recommendations for future work below:

- Identifying more keywords for analysis of decisions.
- Constructing a prediction model for the decisions on applications.
- Framing of insights for both Motions and Appeals together are some directions we can work on.

This CSV File (only motion) included all the data processed from motion applications. This CSV File (motion and appeal) included all the data processed from all applications. This Folder included all the motion pdfs files from LEXIS database which was directly provided by the client. This is the repo to the homepage of this project.

Acknowledgments

We would like to thank Julie Ann Dahlstrom for the opportunity of working alongside her on this project and for her feedback. We would also like to thank Shubhangi Jain and Shruti Gupta for their consistent support, their feedback, and their enthusiasm throughout the length of the project.

Appendix

[1] In the helper function extract_single_pdf_info(), this function extracts the required text information from an individual application PDF:

1. Extract ID or Case number: extraction from the top of page (In Re: 9435010 information) or extraction from the end of page (ID# 1940904 information)
2. Extract Date: extraction from the head of the page: Date: (MAY 23, 2013), Administrative Appeals Office\n() from the next line, or Decision of the Board of Immigration Appeals
3. Extract Decision: search for any texts include appeal or motion
4. Extract Status: search for ('FORM [i1l]-914.*?status', 'APPLICATION:(.*?Status)', 'PETITION:(.*?Status)'); these key words cover all the files which include status
5. Extract Order: search for 'ORDER.*?:.*?(The.*?)\.'
6. Extract Description: search for '(The Applicant.*?)(\s+I\.|$)'; description is very clear and straightforward
7. Extract is_family from files: if the texts include key words like family member, we can tell if the case is a family case
8. Extract LEXIS Citation: every files have this citation, so we use this field as our id to distinguish different rows
9. Extract Content from analysis: based on the texts from the description column, we used a nlp library spacy; by using this library we can easily lemmatize these texts (remove adjectives and remove different tense). In addition, we removed some words from texts such as applicant, law, and application etc. These words have not much meaning in the application. Since some sections of the pdf files do not have an analysis section, we defaultly used description.

* Some additional work also helps us analyze data like cleaning date format.

* Type processing: since our data for type is not binary type (we decide to extract motion/appeal from these texts)

* Description processing: get rid of unnecessary information

* There are three types of order in the application. Motion cases have this information at the end of the application. Thus, we can simply extract this information by using regular expressions. There are these types of order:

- ***The motion is granted***
- ***The motion to reconsider is denied***
- ***The motion to reopen is denied***
- ***The motion to reconsider is dismissed***

[2] This table describes the fields of the dataset, along with possible values for some categorical columns.

Column Name	**Meaning of each column**
dataset_num	Number identifier for files within our dataset.
url	Hyperlink from the title of each pdf file.
file_name	Name of the file.
path	Relative path of the pdf file in the Lexis dataset folder.
ID	Case number of each case.
date	Date of the application.
type	Type of application, can be 'appeal' or 'motion'..
status	Type of Form listed in the application.
order	Text under the Order section in the application, i.e. the decision on the application.
is_family	True if it contains “Supplement A, Application for Qualifying Family Member of a T-1 Nonimmigrant” in status, False otherwise.
description	First paragraph of the application under the Opinion or Analysis section.
Lexis Citation	Lexis Citation number of each application (referred to ID).
counsel	'no' if the application does not contain a Counsel section, 'yes' otherwise.
contents	Text under the Analysis section, or the Opinion section if Analysis section not included.