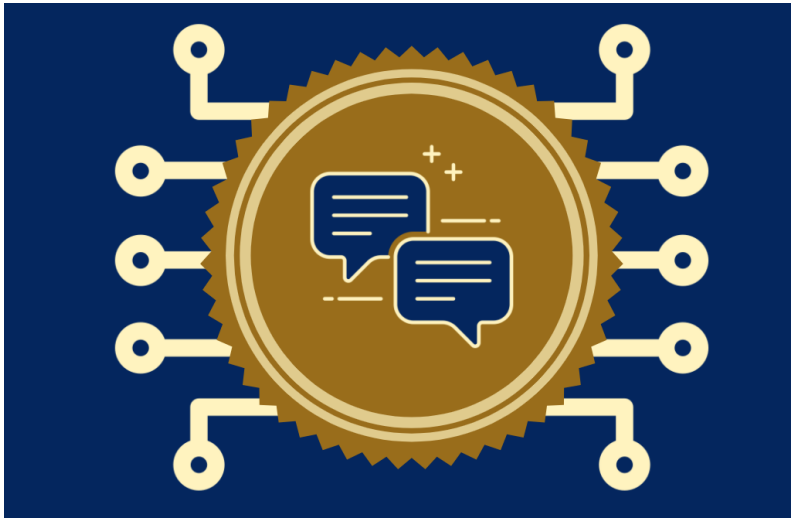
Exploring Passive Data to Synthesize Customer Perceptions for The USPTO



Abstract

The United States Patent and Trademark Office (USPTO) uses surveys to monitor quality of experience. Beyond surveys, an abundance of actionable data exists on the internet. Locating these data requires studying alternate methods to gain insight on customer perceptions. The goal of our project was to study one such method, Passive Data Collection (PDC). PDC uses software to collect data without a customer's explicit permission. These data include social media posts, blog posts and discussion websites. Our research included a literature review of PDC techniques and software, employee interviews, and a SWOT analysis of the USPTO's current system. Based on this research, the USPTO should implement a PDC system and also increase social media use to facilitate feedback.

Team Members

Lillian Garfinkel

Joseph Scheufele

Benjamin Staw

Wayde Whichard

Advisors

Professor Holly K. Ault

Professor James P. Hanlan

Sponsor

UNITED STATES PATENT AND TRADEMARK OFFICE

uspto

B term

December 11, 2020

An Interactive Qualifying Project submitted to the Faculty of Worcester Polytechnic Institute in partial fulfillment of the requirements for the Degree of Bachelor of Science

WPI

Passive Data Collection: A Supplemental Method for Collecting Data on Customer Perceptions

The United States Patent and Trademark Office (USPTO) is the government agency responsible for examining and granting patents. The USPTO's Office of Patent Quality Assurance (OPQA) works to ensure a high standard of excellence throughout the complex patent application process. As shown in Figure 1, the number of patent applications the office receives is growing at an exponential rate[1]. Despite this rapid growth in participation, the patent application process remains a rigorous task. It is so rigorous that 97% of applicants hire trained patent attorneys to manage their application[2]. As the number of applications grows, the number of customer perceptions grows as well. To maintain a standard of quality, the OPQA must always have the best information on the state of its customers' perceptions toward the patent application process.

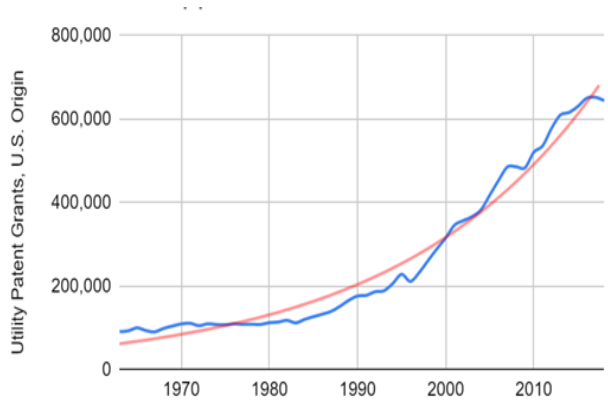


Figure 1: Total Patent Applications

A graph showing the growing number of patent applications since around 1970.

The OPQA measures these perceptions by conducting surveys and administering questionnaires. Figure 2 shows that, since 2013, the rates of good or excellent perceived quality have plateaued between 50-60%, while the rates of poor or very poor perceived quality have trended lower. Despite the decline in poor ratings, the plateau of positive perceived quality was unsatisfactory for the OPQA. More information was needed about customer perceptions to improve the positive perceived quality and not stagnate. Therefore, the OPQA was looking to employ new initiatives[3].

One way to better understand perceptions is to learn more about the customer's journey. This journey spans preliminary patent research, the patent application process and the period after a patent is granted[4]. The internet provides a medium for the USPTO's customers to communicate at any point in their journey from application to approval. There is a gap in knowledge about what these customers are saying on the internet. Therefore, a method of gathering and analyzing these comments and blogs can provide insights into new trends and issues with the services provided by the Office. These issues and trends, once better understood, can be addressed internally and PTO can then improve customer perceptions.

Passive data collection is the study of collecting publicly available data on the internet autonomously, without a user's knowledge or consent[5]. Some advantages an automated system has over surveys and questionnaires include the return of all responses at once, and the ability to search thousands of online locations in seconds[6].

To assess how passive data collection should be adopted at the USPTO, the team established two objectives:

1. Investigate how the USPTO can use passive data collection to understand customer perceptions;.

2. Determine the most ethical, efficient, and informative way of implementing a passive data collection system.

Creating an effective system to collect data from the USPTO's customers required knowledge about intellectual property. It required research about the current data collection system, and how to measure customer perceptions. The scope of this research included examining what others have done for passive data collection and how to analyze the collected data.

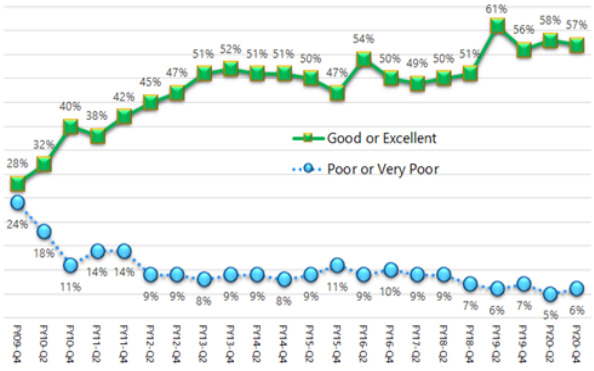


Figure 2: Bi-quarterly Customer Quality Ratings

A graph showing a rise of customers rating the quality of the USPTO as good or excellent, and a fall of customers rating the quality as poor or very poor. The data is from 2009 to 2020.

Patent Applications: A Challenge for Inventors

Intellectual property is any idea that can be created through artwork, invention, design, or symbols[7]. Global standards protect the rights of creators and owners of intellectual property. This ensures creators benefit from their own work or investment. Protections encourage creators to produce more work, creating new industries and growth in the economy. Ultimately, these attributes lead to increasing the quality of life in a society. The most common form of protecting intellectual property comes in the form of a patent[7].

Patent applications require a detailed description of the invention. The USPTO states that "while a patent may be obtained in many cases by persons not skilled in this work, there would be no assurance that the patent obtained would adequately protect the particular invention"[8]. If a patent is not filed properly and thoroughly reviewed, its protections could be unenforceable. As a result, nearly all prospective patent applicants hire patent attorneys to help with their applications[2].

To acquire a patent, an applicant must follow the USPTO's guidelines. The applicant must ensure that their invention has not already been patented. If the invention is novel, the application must be of the correct type. After filing, the application is reviewed by a patent examiner in the appropriate field. If approved, a patent is granted following payment of the issue fee and the publication fee. For utility patents, a maintenance fee must be paid 3-3.5 years, 7-7.5 years, and 11-11.5 years after the date of issue[9]. Located in the supplementary files is a flowchart detailing the patent application process. As customers file, touch points (interactions between examiner and patent filer) are critical in making sure a patent is fileable[2]. Based on the diagram, there are only three touch points between start and finish:

1. Patent application is filed;
2. Patent examiner issues an office action;
3. Associated fees are paid.

The patent application steps are complex and difficult for the average customer to navigate alone. Due to the large volume of applications, patent examiners work with multiple customers at once[1]. There often is not enough time to discuss feedback outside of the customer's technical work. The lack of touchpoints throughout the application process has compelled the USPTO to look for alternative settings to collect data on its customers' perceptions.

Axioms of Measuring Customer Perceptions

Customer outreach is an essential practice to any organization or company that provides a service[10]. With the increase in use of social media, customers feel that reviews left by other individuals on company websites are more reliable than are advertisements from the companies themselves[10]. As a result, many companies monitor social media sites. Comments left on company pages on sites like Facebook, Twitter, and other third-party blog pages have proven to be valuable sources of customer data. Interpreting online comments will allow the USPTO to better understand its customers' perceptions and form a stronger connection with the views of its customers. Therefore, the USPTO has an interest in developing a more in-depth system to measure customer perceptions.

Customer perceptions are derived from sentiment, satisfaction, and quality of experience (QoE). Customer sentiment reflects the emotion customers feel toward a certain service over a period. Sentiment cannot be measured quantitatively and requires context based on time, and type of service[11]. In contrast, customer satisfaction measures how a customer feels about a service relative to a particular moment and can be measured quantitatively. Satisfaction is generally split into 3 levels: poor, fair and good[11]. Very poor and very good can also be included. QoE is the quantitative rating of an overall experience with a service. A metaphor to assist in understanding customer perceptions goes like this: if customer perceptions were graphed with time as the x axis and perception as the y axis, then sentiment would be a continuous line, satisfaction would be the slope of the line at a particular point, and quality of experience would be the area under the line.

Commonalities between sentiment, satisfaction, and QoE are analyzed to interpret customer perceptions, compared to customers' experiences. The Expectation Confirmation Theory (ECT) shows how this can be useful[12]. For example, when a customer is considering a purchase, they have an expectation of how the product will perform. That product's performance could either match that expectation or not. ECT studies the difference between a customer's expectation of a service and its reality[12]. The USPTO could use ECT to analyze the relationship between customer sentiment and satisfaction to gain an understanding of customer perception. Figure 3 is a graphic of how satisfaction is defined based on the Expectation Confirmation Theory. However, ECT can only be effective when there is plenty of feedback

spanning all aspects of the process.

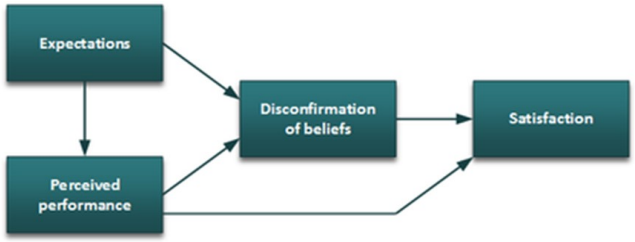


Figure 3: Expectation Confirmation Theory Diagram

A diagram explaining the process of Expectation Confirmation Theory (ECT).

Prior Efforts to Gauge Customer Perception at the USPTO

The USPTO does not receive feedback at enough points in the patent application process. For over 20 years, to supplement the lack of feedback, the USPTO has solely conducted surveys and questionnaires in exploring its customers' perceptions of the application process[2]. These surveys and questionnaires fail to reach an accurate representation of the full customer population, because people must choose to participate.

Due to the nature of the USPTO's surveys and questionnaires, a void of knowledge exists regarding customer perceptions. In recent years, the results of these surveys and questionnaires have plateaued at around 57% approval and 9% disapproval[2]. The USPTO has recognized that measuring and understanding customer perceptions will no longer improve using only these methods. One proposed cause for this is that surveys and questionnaires do not encompass full representation of non-frequent filers[13]. When these people are not reached out to, they are left unheard. Therefore, it may be fruitful to see if these same people are talking about their experiences elsewhere than at sites usually accessed by PTO on the internet.

Collecting Customer Perception Data

Organizations often collect data from their customers to enhance their understanding of the customers' experience. In addition to traditional ways of collecting this data, like surveys and questionnaires, organizations have begun to look for customer perceptions online[14]. This is typically done without direct involvement from the customer and is known as passive data. Passive data collection is a method of collecting data without explicit contact with a customer[5]. With passive data, an organization can create a unique database of customer data from specifically chosen online sources. These sources may include websites or social media pages with comment forums related to the business of the organization. Passive data collection invites feedback from customers who may not be offered formal surveys or are reluctant to take time to respond to them. Therefore, collecting passive data offers an opportunity to broaden an organization's customer sentiment database.

While collecting data passively is effective, an improperly structured collection system could lead to unintended consequences. Topics to consider include data privacy, data credibility, source credibility, and the effects of the program itself. For example, in 2014, QVC Inc. sued Resultly LLC[8]. The web crawler Resultly deployed on QVC's website was too aggressive and shut down QVC's website for a period. As a result, QVC lost substantial sales and was unable to provide services to its customers. Due to these events, the team's research needed to comprehensively investigate the unintended side effects of any possible strategy the team planned to recommend to the USPTO.

An important factor to consider was how to approach collecting customer perception on the internet. There are various components to passive data collection, the first being the collection method.

Every website on the internet uses a coding language called Hypertext Markup Language (HTML). Figure 4 depicts the organization of HTML code from the USPTO's website. Inside the first two red circles are common labels used by HTML to organize code like a tree. As a result, each 'parent' section has 'children.' This structure allows a tool to search quickly through the code. The last three circled sections depict text that is visible to users. These sections and others contain all the necessary information that makes the USPTO's website viewable. HTML has sections for links connecting websites, images on the page, text, animations, and advertisements[15].

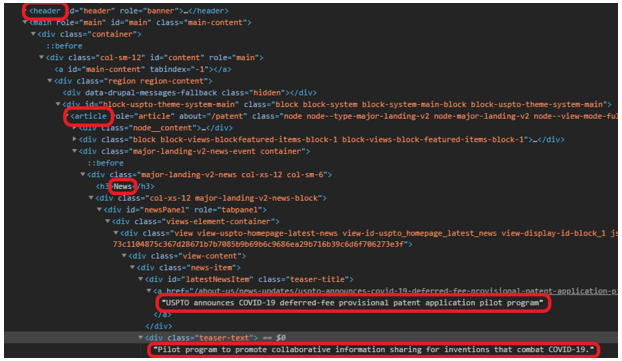


Figure 4: HTML code from the USPTO Website

A snippet of code from the USPTO website for reference of HTML structure.

Web scraping takes advantage of the HTML structure behind websites. A web scraper can visit thousands of web pages and interact with them like a human can. Among other actions, a web scraper can fill out forms, and read text[10]. Since a web scraper does not open these pages the same way humans do, it does not have to wait for any pages to render. It can also completely scan or interact with multiple pages in milliseconds. This device is ideal for collecting information from multiple sources such as spreadsheets, survey responses, and commentary from forums.

Web crawling is another form of passive data collection. A web crawling program moves from website to website autonomously. It searches for specific identifiers provided to it by the organization using the web crawler. These identifiers consist of a dictionary of words related to customer perceptions. Once the identifiers are found, the program will create a sentiment level analysis score[16].

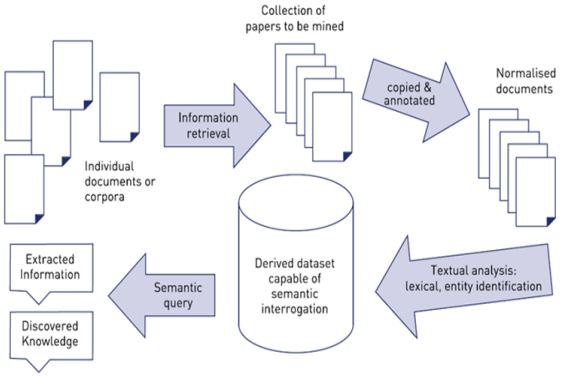


Figure 5: Flow Chart of Text Mined Data

A flow chart showing the process for text mining data. The data is inputted to the text miner and useful information is returned using semantic analysis.

Data mining can collect large swaths of data from websites specified by the system administrator. Generally, this data all appears in the same format. For example, Lyu and Choi employed a data mining system to collect product reviews, price discounts, number of reviews, and organic labeling[17]. Text mining differs from data mining because it considers text only written by a user. Text mining traverses text documents on webpages to analyze words and assign them to specific feelings, emotions, or level of quality. Text mining software would allow for many responses to be quantified. Figure 5 is a flowchart to illustrate the use of text mined data.

Figure 5 displays the process by which researchers from the University of Cambridge analyzed their documents with text mining. Following the arrows around, the initial step of the process is to identify documents for analysis. Second, the documents are then converted into machine readable format (text and code, no pictures). Third, the text mining software scans the document. It provides any kind of analysis as specified by the software administrator. Finally, the extracted information then becomes discovered knowledge for the researchers to use.

Analyzing Customer Perception Data

The next step after collecting the data is analyzing it. Data analysis allows an organization to draw accurate conclusions about its customers' perceptions. When considering sentiment analysis, there exist two approaches, machine learning and lexicon based. Figure 6 shows how each technique of passive data analysis is categorized.

Machine learning is a state-of-the-art method of data analysis and excels at identifying patterns

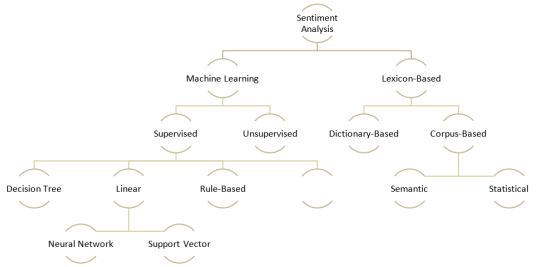


Figure 6: Hierarchy of Sentiment Analysis[5]

A flow chart showing the process for text mining data. The data is inputted to the text miner and useful information is returned using semantic analysis.

and making predictions[7]. Machine learning is based on the idea of creating neural networks, essentially large linear algebraic equations. Neural networks can be used to analyze the intent of sentences. This method of machine learning analysis is called natural language processing[18]. Sentences can be passed through the neural network, returning a set of outputs that categorize each input sentence.

A neural network has two phases, a training phase, and a predicting phase. During the training phase, data with a known category is provided to the network. Then the network attempts to predict the category. Based on the accuracy of the prediction, a small correction is made to the network that ideally produces a better result. After many iterations, the network will learn to make better predictions without memorizing the training data set. After training is complete, the prediction phase begins. In the prediction phase, the neural network takes in data it has not seen before and outputs a categorical prediction[18].

Passive Data: The Path to a Solution

Preliminary research indicates that there is a need at the USPTO for customer perception analysis. This analysis can lead to advancing the customer perceptions in ways that the current system cannot. Thus, the team proposed to the USPTO a recommendation for a passive data collection plan. The methodological approach the team took is described below.

Exploring a Passive Data Collection Strategy for the USPTO

This methodology addresses two research questions. The first question is how the USPTO can use passive data collection to understand customer perceptions. The second question is what would be the most ethical, efficient, and informative way of implementing a passive data collection system.

To answer the first question, the team:

1. **Conducted** a literature review of passive data collection plans;
2. **Identified** information the USPTO can act on, for the purposes of understanding customer perceptions;
3. **Reviewed** the USPTO's current understanding of customer perceptions;
4. **Selected** sources from which data could be collected.

To answer the second question, the team:

1. **Performed** a SWOT (Strength, Weakness, Opportunity, Threat) analysis on the USPTO's current data collection system;
2. **Reviewed** the limitations of each possible strategy;
3. **Created** a decision matrix to determine an optimal passive data collection system.

The team accomplished these tasks using the following methodologies: literature reviews, interviews, content analysis, decision matrix, and SWOT analysis. The literature review analyzed current passive data collection systems in use by professionals and other organizations. In detail, the literature review showed what aspects of these systems could help the USPTO. Interviews elicited information from experts in information technology and customer service. Interviews also elucidated the team on how the current data collection system operates. The content analysis aggregated responses from the interviews to pinpoint the most frequently occurring topics of highest priority. A SWOT analysis highlighted the strengths, weaknesses, opportunities, and threats of the current system of surveys and questionnaires. The team created a decision matrix to organize components of passive data collection practices. Figure 7 is a chart that visualizes the team's methods.

Question 1: How can the USPTO use passive data collection to understand customer perceptions?

Due to the stagnation in knowledge gained from surveys and questionnaires, the USPTO was interested in using passive data collection to better understand customer perceptions. The USPTO was not fully cognizant of many techniques involving passive data collection. However, passive data collection provides service-based companies a much larger and less biased dataset of opinions[19]. To recommend a possible approach, the team conducted a literature review of passive data collection practices.

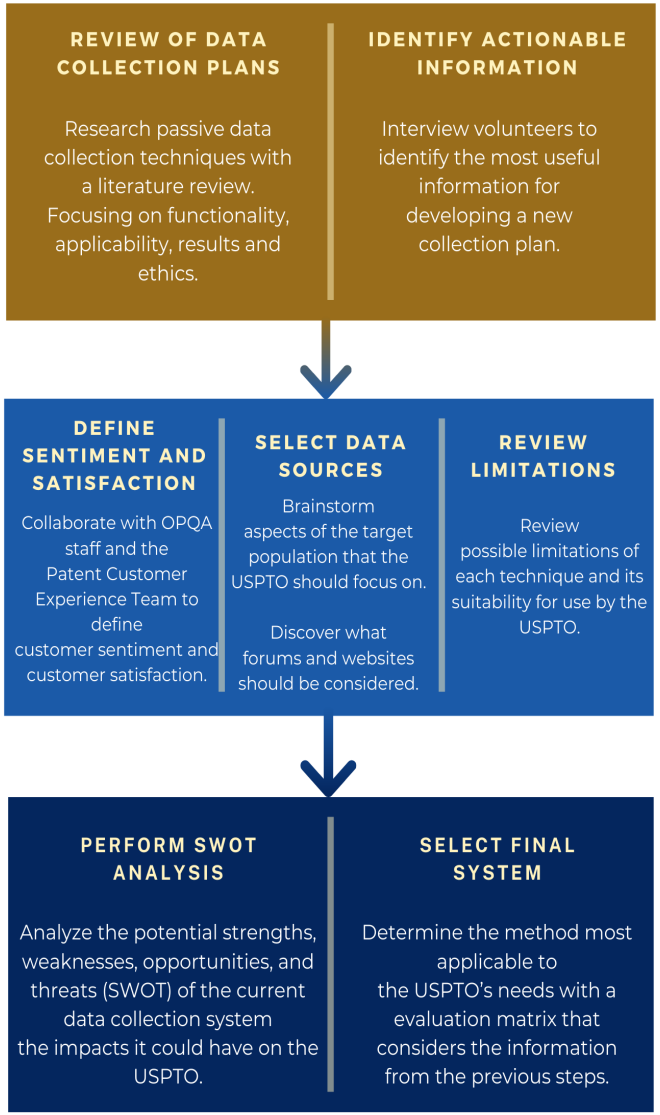


Figure 7: Methods Flow Chart

A flow chart of methods presented in sequential order. Each box includes the scope of each research objective and direction of analysis.

Review of Data Collection Plans

The literature review was broken into two categories: data collection and data analysis. In both categories, the team conducted a review of several techniques. These techniques can be seen in Figure 8 . In addition, each source was probed for: functionality, applicability, results, limitations, and ethical implications.

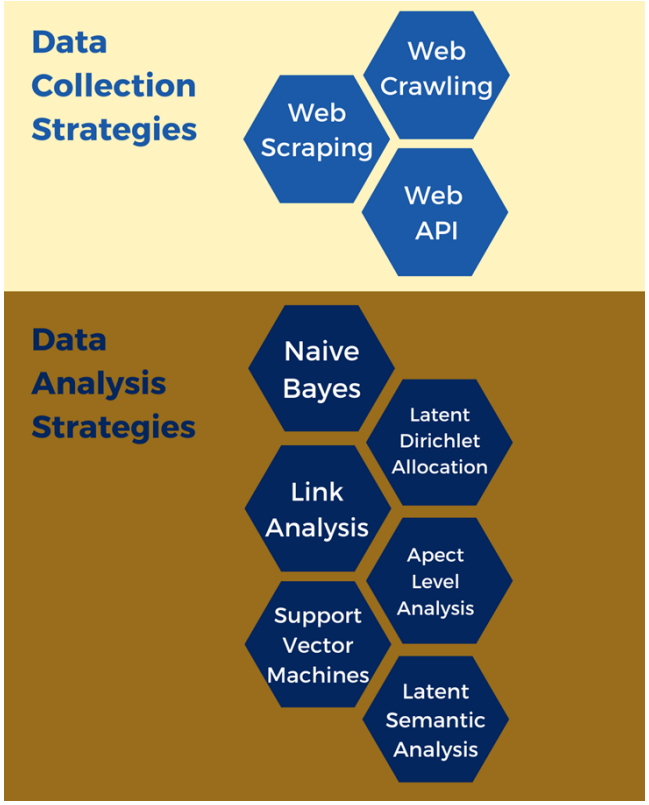


Figure 8: Literature Review Topics

A graphic depicting the topics covered in the literature review found in the supplementary materials.

Identify Actionable Information

Passive data collection provides actionable information for the USPTO. For instance, the USPTO wanted to have a better understanding of how customers view the patent application process. In theory web scrapers could scrape comments on a patent blog about fee increases, for example. These comments might raise a flag for the USPTO to decide whether to act on this information. The purpose of understanding actionable information was to define the elements of customer perception[20]. Then, the elements were used for the passive data collection strategy. To define what information the USPTO can act on, interviews were conducted with the Patent Customer Experience Team (PCET) and several staff members from the OPQA.

The information gathered from interviews defined the specific types of actionable information that guided research objectives. The team was able to pinpoint specific strategies to measure customer perceptions. To determine if passive data collection could inform the USPTO, customer perceptions need to be defined. The main factors that directly affected sentiment at the USPTO included, but were not limited to, timeliness, accessibility to applications, and receipt of patents. The USPTO determined customer satisfaction by analyzing their sentiment[4]. Figure 9 is an example of the USPTO's perceived changes in quality from 2018 to 2019. The figure shows a relationship between how the customers' perceptions changed. To define the factors that affect customer perceptions of quality, the team interviewed WPI faculty, students, and patent attorneys. All the participants had experience with the application process.

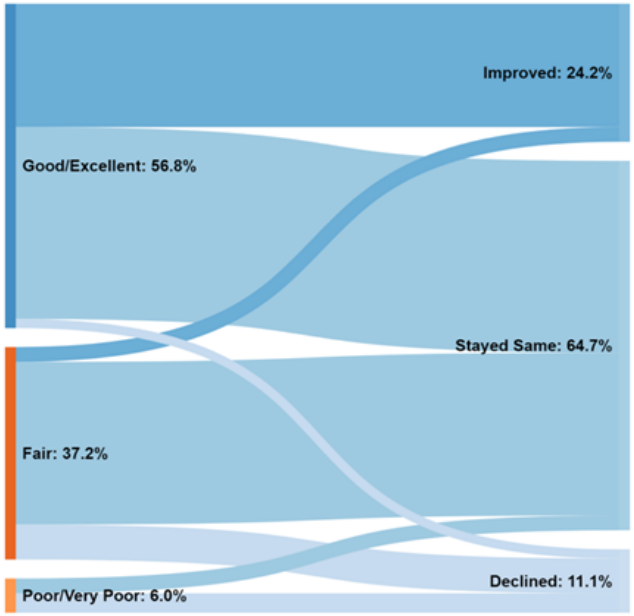


Figure 9: Perceived Quality of the USPTO

A flow chart showing the process for text mining data. The data is inputted to the text miner and useful information is returned using semantic analysis.

Select Data Sources

Any potential passive data collection strategy needs to be tailored to the USPTO's customer population. Thus, the team considered a range of web addresses from which information and data can be collected. To assist in finding websites, interviews were conducted with IP specialists and members of the OPQA. Technology commercialization professionals at WPI and the OPQA are well qualified to speak on customer relations. Together they informed the team about websites and web forums that can provide insight into customer perceptions from customers of the USPTO.

Question 2: What would be the most ethical, efficient, and informative way of implementing a passive data collection system?

To understand the answer to this question, a SWOT analysis on the current data collection system was conducted. A review of the limitations of the current data systems was also needed. A design matrix aggregated all the aspects surrounding the techniques. The matrix allowed the team to select an appropriate passive data collection system.

Review Limitations of Passive Data Collection Plans

After research was conducted into the number of websites being targeted, the collection and analysis techniques, and what information the USPTO can act on, the team considered possible limitations. A literature review of data analysis strategies and interviews with USPTO employees revealed limitations of each strategy. Each plan had different requirements such as build time, efficiency, and quality of software. These factors were compared before making a recommendation to ensure the team presented the best solution.

Select the Final Passive Data Collection System

To provide the USPTO with a passive data collection plan, the team needed to determine methods to collect and analyze data. The techniques discussed in the team's review of practices were considered, including both collection and analysis strategies. A decision matrix was created to weigh the differences in using each analysis technique. A separate paper was written detailing the explanations for each score. All this information can be found in supplementary files.

Perform a SWOT Analysis

To understand where passive data collection can fit into the USPTO's current system, the team performed a SWOT analysis. This type of analysis was used to determine the strengths, weaknesses, opportunities, and threats of the current data collection system at the USPTO[22]. Figure 10 depicts a matrix for understanding the relationships between the four criteria.

Each square of the matrix was filled in with assistance from members of the OPQA, IT and others. Strengths were defined as areas within the USPTO where the current methods succeed. Weaknesses were defined as areas within the USPTO that need improvement or require assistance. Opportunities were defined as areas that the USPTO can leverage. Threats were areas that could cause harm to or create issues for the USPTO. The team conducted interviews with the USPTO Staff members to understand their opinions of the strengths, weaknesses, opportunities, and threats of the current data collection system.

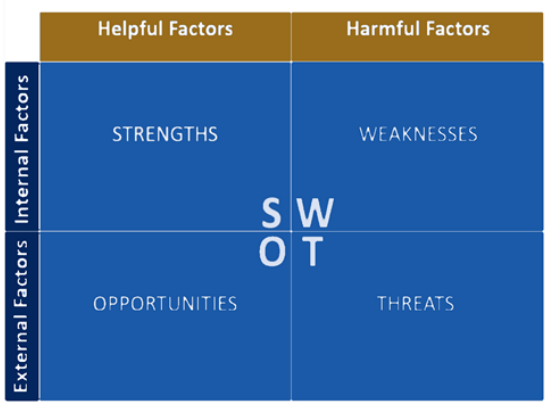


Figure 10: SWOT Analysis Example

An outline of a SWOT analysis.

Practices of Passive Data Collection and Analysis

This section will discuss the team's findings, and how they relate to the USPTO. The main portion of the team's research consisted of a literature review of practices in passive data collection and analysis. These findings include a discussion of:

1. What each technique is;
2. How each technique works;
3. How the various techniques can be used to understand customer perceptions;
4. How open-source software can be used to automate the process.

Web Scraping

Web scraping is the process of collecting and storing large amounts of data from the web for further analysis. Web scraping requires the use of software. There are many different forms of software, each involving differing levels of human interaction, programming expertise and ease of use.

The first kind of web scraping tools are a family of web extensions. A web extension is an application that mounts onto a user's web browser, adding functionality[21]. Web scraping extensions allow the user to extract data with very few clicks. One example of a web extension designed for web scraping is AnyPicker. AnyPicker is available on the Google Chrome web store. When AnyPicker is turned on, the user will click and drag a box over the areas from which they want to collect data, as seen in Figure 11.

The grey boxes on the right represent html elements with their text data. On the left of the screen is the set up for the web scraper. The extraction rules essentially lay out what the scraper will collect. Next the data source list must be set. Most web extension web scrapers rely on the code being standardized (i.e. website structure is predictable). Due to the tree-like nature of HTML, a list of pages from the same website will be identical in structure but differ in content. Finally, the web scraper runs and collects the specified information from all the data sources specified and places them into a .csv file. An additional feature included with AnyPicker is a timer and request limiter to control how often the scraper accesses each web server[23].

When searching for 'scraper' in the Google Chrome web store, many products appear. All of them will operate in the same way as AnyPicker. The key areas to look for when evaluating a web extension web scraper is whether the creator limits the number of pages that can be scraped per month, the availability of free-trial periods, the quality of user interface and the ability to save the data to a database or spreadsheet[24].

Another type of web scraping involves creating a native application in Python using a Python package. One major advantage to building a native app is that it allows for a streamlined system. It allows one program to web crawl, web scrape and text mine without having a human intervene to run separate programs. A web scraping Python app would be one part of this program. Another advantage is that Python allows for upgrades or changes to the system. A web scraping Python package can be found on pypi.org, which is a repository for open source projects to be shared in a format that can be easily installed and added to any Python program[25]. The source code for these packages can be found on the developer's GitHub (a community for code sharing)[26] or in the documentation supplied to pypi.org[25]. The two most popular web scraping packages available include Scrapy, and Beautiful Soup[27]

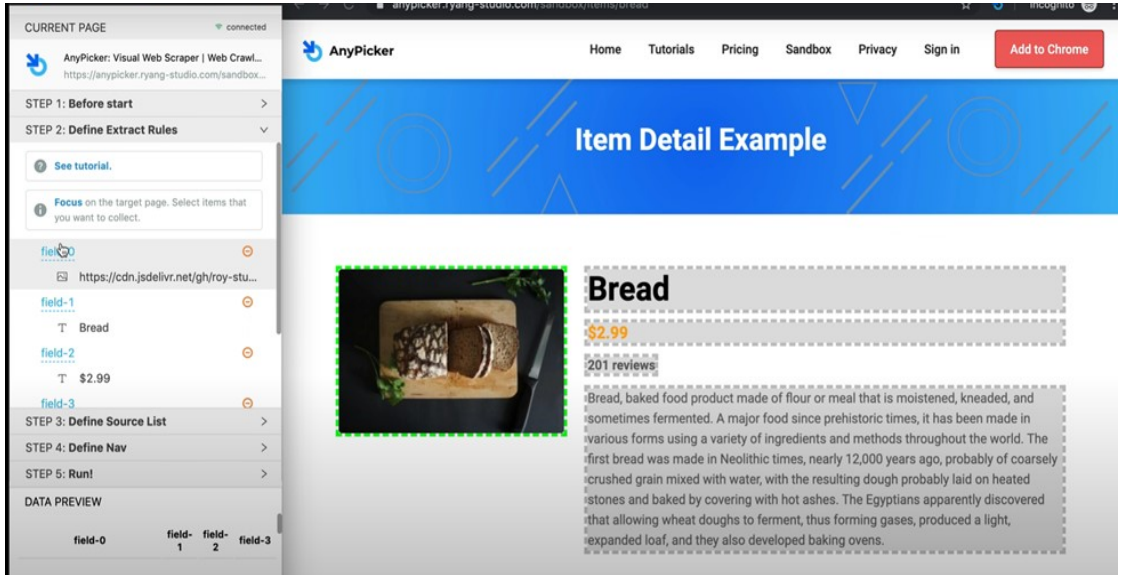


Figure 11: AnyPicker Example

The picture shows what it looks like to use AnyPicker to web scrape a website.

Web Crawling

Sometimes organizations require many different perspectives from many locations to ascertain customer sentiment, satisfaction, and QoE. This can be achieved using a web crawler. Starting at an initial specified URL, the web crawler will traverse the whole webpage for other related URLs. It indexes the information it finds at each link, then proceeds to other related pages to do the same. The crawler itself does not analyze any information, it just constructs the path of pages for another method to collect and then analyze[28]. In one study, a web crawler was deployed that started off looking at TripAdvisor for reviews, then branched off to other hotel websites from there. This gave the researchers insight on customers views of a provided hotel service[29].

Web APIs

Organizations may want to collect a large amount of data from a few select websites. When conditions are appropriate, this process can be expedited with the use of an application programming interface (API). APIs are software written to assist in a larger program. They are written to allow for more abstraction. APIs can be disseminated to the public to reduce redundancy[30].

To use an API effectively, the user must be able to work with higher level programming languages such as JavaScript or Python. APIs written in this code are usually broken up into browser APIs and third-party APIs. A browser API is built into the web browser and can alter, collect, or organize information. Alternatively, a third-party API requires an additional application, website, or software to perform those tasks[31].

Once code is written to direct an API on actionable information, it can help researchers find trends in data. For example, an API can gather data from a server to update a user about new information. If a user sets an API to notify them every time someone posts on Twitter about intellectual property, the web API will regularly query the server until it finds text related to intellectual property.

Latent Semantic Analysis

Latent Semantic Analysis (LSA) is an analysis method used for finding links between words without their explicit definition by contextualizing them[32]. It is capable of simulating human phenomena such as learning vocabulary words, word-categorizing, recognizing words derived from others, understanding conversations, and judgements of essay quality[33]. Landauer and his team studied LSA and tested how well it can distinguish words relative to a human. When prompted to select the best synonym to a given word, out of 4 choices, LSA was 65% accurate. This result came after training the system with over 4.5 million words from 30,000 encyclopedia articles[33].

The first step in LSA is to represent the text as a matrix where each row stands for a unique word, and each column stands for a text passage. Then, once the matrix is created, singular value decomposition (SVD) is applied to the matrix. SVD is a form of mathematical generalization that goes beyond the subject matter. The product of conducting SVD is a least-squares best fit of the frequencies. A least-squares best fit finds the line that crosses the most frequency points, Figure 12 is an example of this.

The dots represent data points, and the line is an estimation of the frequency. LSA can predict how often words can appear in a text, even when they have not appeared. That fact can be used to find the themes of what is being written[33].

Latent Semantic Analysis models can be created with the fitlsa function in MATLAB[34]. MATLAB is a trusted industry standard; it is used by million-dollar corporations and universities alike. All that is required for inputs is the matrix of frequencies, and the number of components. What results is the least squares best fit matrix[34]. LSA can also be conducted directly using Python, with Python's default Math package. A function can be created that is like fitlsa with the same inputs and outputs, however it would require more effort from the programmer[27].

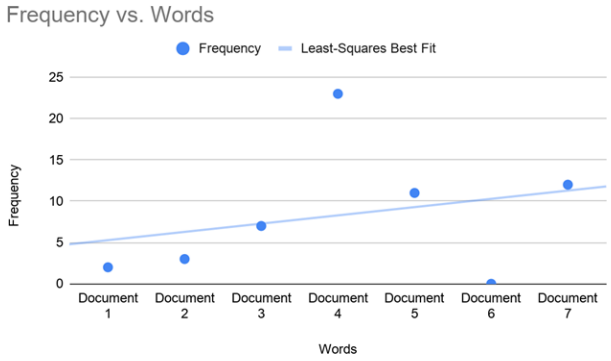


Figure 12: Linear Regression Example

An example of a Linear Regression of the frequencies of the word USPTO over 7 documents.

Latent Semantic Analysis can provide the USPTO with topics that customers are talking about in their survey responses and any other repository of customer feedback. These topics would be generated by considering the context of the surrounding words. Since some forms of LSA require training, and others do not[32,33], LSA may not require as much set up time, but it can still produce similar results to other methods.

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a probabilistic topic modeling technique. An LDA algorithm takes in a set of word-containing documents. Each document is modeled as a certain combination of topics, and each topic is modeled as a certain combination of words within the documents[35]. Associating words with topics makes trends and commonalities between documents more apparent. The probabilities of topics among documents are then calculated to provide a representation of the contents of the documents[36].

In addition to finding trends, there are different styles of data analysis using LDA. An LDA model is capable of accurately predicting whether a new comment belongs to a certain topic and whether it is positive or negative once the frequencies of the words are determined and the model is trained with the appropriate data[17]. Lyu and Choi believe LDA is exceptional at procuring topic and subject words from tomer's sentiment. The dimensions are determined based on the probability distribution. However, before using LDA, the textual data needed to be

cleaned for meaningless words, which was time consuming. Separate programs were written purely for the preprocessing of the data37. If this style of LDA can be used to gather the factors of customer sentiment, then a second search could be conducted studying the conversations around the same factors.

One major reason for the USPTO to collect customer data is being able to predict trends. LDA can discover hidden trends in big data[35]. If the USPTO were to use a method like LDA, trends in data would be easier to find. Additionally, trends could then be analyzed to understand the expectations regarding the application process.

Link Analysis

Link analysis is used to find relationships between data. Links are found by identifying commonalities between two sets of data. Frequently, link analysis is used by law enforcement to understand the connections between criminal networks[38]. Most link analysis strategies were specifically used to quantify data and correlate relationships between trends, requests, and actionable information.

To conduct link analysis, multiple data sources are required. Some example inputs of data sources are events, websites, interviews, and people[38]. The information gathered from each source must be recorded. For example, important information from an interview will not be collected unless interviewees' responses are recorded. Then, further analysis of the information gathered is required. Finally, according to Andrew Disney, lines can be drawn between each source, or the information from each source. When lines are drawn, they can be based on similarities, differences, or trends[39].

Link analysis at the USPTO can help sort actionable information from multiple data sources. Posts that are made on Patently-O, IPwatchdog, Reddit, Twitter, and Facebook can all be collected. Link analysis can then be applied to compare the trends, similarities, and differences of information from each source. For example, if one person posts that the timeliness of the application process is too long, there can be multiple comments in response to this post. The entire chain of responses is called a conversation. Throughout this conversation there may be many forms of actionable information that can lead to deeper sentiment analysis. When similar conversations happen on a different source, the utilization of link analysis will allow a link to be drawn between the similar information from each source. Overall, this leads to analysis of common trends in the actionable information.

Support Vector Machines

Support Vector Machines (SVMs) are a type of neural network that can be trained and programmed to classify textual inputs into categories. When applied to a large data set, these classifications can establish the general topics or emotions of what is being said[40]. Grljevic and Bosnjak found that their SVM algorithm was about 80% effective in predicting negative comments and about 76% effective in predicting positive comments taken from a variety of review websites, like IMDB and Yelp. In all trials, the algorithm was more successful at identifying negative connotations than positive[41]

Support Vector Machines are designed to find the hyperplane of best fit. The first step is to translate the text data into frequency data. The hyperplane results in a boundary between the frequency data points. The size and shape of the boundary depends on the dimension. The boundary created by the hyper plane is called the decision boundary[42]. Any input will be classified by its position relative to the boundary. For example, in 2-D a hyperplane is a line. If an input value falls above the line, it is classified to one category. If it falls below the line, then it is classified to the other category[42,43]. The 2-D hyperplane does not need to be a straight line, it can be circular, quadratic, or exponential in shape. Support vector machines are largely controlled by their Kernel Function. The kernel function is a starting point for the system to converge to the optimal hyperplane shape. If the USPTO were to use an SVM, the kernel function can be determined by looking at the collected and plotted frequency data and using trial and error.

One open source library of support vector machine software is TensorFlow. TensorFlow is a project created by Mozilla. It houses an API that can create an SVM. The requirements include the input dimension (based on number of satisfaction factors), kernel function, training data and prediction data. Some programs can be pre-trained, or training data can be found. TensorFlow is an industry standard used by Twitter, GE, Coca-Cola and Google[44].

Another open source library for SVMs is LIBSVM. If all that is required is to build an SVM with the most ease of use, then LIBSVM is a great choice. It does not have the additional unnecessary features of TensorFlow and is therefore contained in a smaller file and may run faster. LIBSVM uses a cache to store results from previous iterations of the SVM. Torres-Boran used LIBSVM to standardize a testing environment for studying different SVM kernel

functions on the same set of data[45].

Support Vector Machines pose the ability to extract the popular categories from a collection of text. These categories would include keywords related to customer sentiment and satisfaction. It could also be used to classify certain comments as being related to customer perceptions. Building a support vector machine for the purposes of this project may be unnecessary and open-source software may be pursued instead.

Naive Bayes Classifiers

A Naive Bayes Classifier is a classification algorithm that incorporates Bayes' Theorem[46]. Bayes' Theorem asserts the probability of an event, provided another specific event happened as well. When the algorithm is given training data, Bayes' Theorem is used to make a prediction to improve the algorithm's understanding of that data. For example, Mushtaq and their team investigated how factors (such as terminal and network types) contribute to the QoE of video streaming delivery over cloud networks. This was accomplished by employing a Naive Bayes Classifier to sort the factors into categories to be analyzed individually[47]. This technique would have to be implemented alongside other techniques, as it cannot parse text and determine sentiment in an efficient way on its own. For the USPTO, different factors would be selected to capture QoE of the patent application process. The factors from this source do not affect the way the system operates.

The USPTO could use a Naive Bayes Classifier to determine customer satisfaction in a specific area given a circumstance. Hypothetically, the USPTO could be interested in sorting through data to find only comments of negative sentiment related to a certain part of the application process. Apache Spark is a machine learning tool that can use Naive Bayes Classifiers for predictions. It works by using a network of other computers to compute complex problems[48]. Apache Spark can also be used for document classification, meaning it can categorize documents based on their text[49].

Expectation Confirmation Theory

When a customer is considering purchasing a service, they have an expectation of how they believe the service will perform. If the customer buys the product or service, the product will perform in a way that either matches the customer's expectation or does not. The study of this mindset is called Expectation Confirmation Theory (ECT)[50]. When using ECT, the area of interest is the difference between customer expectation and actual performance[51]. This leads to Expectation Confirmation Theory showing that satisfaction is a result of customer confirmation[52]. Machado utilized ECT to evaluate customer satisfaction with hotel services in Peru. They analyzed survey inquiries, and considered customer reported satisfaction and hotel booking prices for context. Ultimately, ECT made sense, as hotel features were determined to have a direct impact on customer satisfaction[53]. Like Machado, the USPTO can benefit from employing Expectation Confirmation Theory. If the USPTO is cognizant of how its customers perceive the patent application process before and during customers' journeys, comparing each customer's expectation to their final experience could be valuable.

Ethics

As seen above, data collection has the potential to make a significant impact. However, if practiced unethically, it can also cause a significant amount of harm. For example, large scale data collection is currently used to understand credit worthiness of consumers. Even if most consumers are not fully aware of what is happening, some argue that privacy is being infringed[54]. Privacy can also be reduced during counterterrorism efforts when many people's data is collected to learn more about a smaller subsection of criminals[55]. In some cases where there is a large amount of data to be collected and analyzed quickly, that data is collected autonomously. This causes a privacy concern as the autonomous collector makes the decision whether to collect specific data and needs to understand when it should proceed and when it should not[56]. People do not always realize how much data they are inadvertently providing to big corporations. When many people realize how much data they are giving away, they frequently become uncomfortable. This limit is often reached without some people realizing[56]. Given that the USPTO's goal is to use passive data collection to learn about its customers' perceptions with the hopes of improving that perception, it is important that attention is given to the amount of data being passively collected. Given that excessive data collection is perceived negatively, the USPTO can better protect its own image by being intentional about the data it collects.

Understanding the Current System

The team conducted interviews with USPTO employees to understand the current system of customer data collection and analysis. The feedback revealed both positive and negative aspects of the current system. The USPTO conducts new initiatives to stimulate customer interaction with the office[13]. However, an occasional issue with these initiatives is that the USPTO fails to detect when the initiative is ineffective or could do something differently. Another finding was that the USPTO surveys its customers and employees for feedback on its operations. However, the responses rely on those who participate, which may not be a full representation of the population. For example, the USPTO conducts quarterly surveys on their most frequent customers. Some individuals felt that the USPTO only hears from those customers who are very satisfied or very dissatisfied with their experience and misses out on the customers in the middle [13]. Ninety-seven percent of the USPTO's customers are patent attorneys, the other three percent are individuals. Due to the small percentage, these applicants are generally forgotten or unheard.

The USPTO also seems interested in having a better awareness of when there are new problems. Currently, the only ways the USPTO learns of new problems is when contacted directly, or when the issue is included in a survey [13]. When an applicant feels their application has been rejected unfairly, bias resulting from their decision often affects their feedback. If these customers were included in the quarterly surveys, the views of the total population would be more completely encompassed.

Identifying Actionable Information

Based on the results of the interviews, the team found that the USPTO needs to collect more data from its customers[13]. Although the USPTO's current system provided some insight into customer perceptions of the patent application process, there were many apparent gaps in knowledge. For example, some interviewees suggested that finding ways to better include less frequent filers in the patent application process would give the USPTO a more complete understanding of what customers want, as these opinions are less noticeable compared to the frequent filers[13]. Less frequent applicants were both smaller in number and under surveyed. Other interviewees noted that searching blog sites could reveal more perspectives. However, they were concerned about verifying the validity of statements found on the internet to ensure the feedback was actionable. Other participants believed that having a better understanding of customer opinions and likes or dislikes would be helpful. Essentially, if there was a way to get more detail from customers during their experience, there would be more actionable information about how to meet their needs[13]. Figure 13 is a word cloud with examples of keywords for the USPTO to use when searching for customer feedback on the web.

Figure 13: Example of Actionable Information

A visual providing examples of keywords for the USPTO to use when searching for customer feedback on the web.

SWOT Analysis

To guide research, interview responses were coded to understand the strengths, weaknesses, opportunities, and threats (SWOT) of the USPTO's current data collection system. A graphic representing the results of the SWOT analysis can be found in the supplementary materials file.

Strengths

Throughout the interview process several themes appeared relating to the strengths of the current data collection systems. One theme was that surveys and questionnaires provide the USPTO direct interactions with customers. The office will host or attend events where they are able to ascertain the sentiments of their customers[13]. Another theme that prevailed was having the ability to tailor questions. The office can probe for opinions on specific issues and get direct answers. With tailored questions, the USPTO can anticipate biases better. Besides tailoring the questions, the audience can be tailored as well. Specific trusted members of the customer base are reached out to. Many of these strengths pose a problem of blocking outside ideas, thus this constriction presents many weaknesses.

Weaknesses

Interview results revealed multiple weaknesses with the USPTOs current data collection system. For example, responses indicated that much of the population is not represented in the responses. Furthermore, statistics generated from surveys are not necessarily representative of the whole population. Another key weakness to the USPTO's data collection plan was that internal surveys did not reach external customers[13]. Interview participants noted that, since external customers do not receive the internal surveys, their opinions of the patent application process go largely unheard. Many of these issues stem from the lack of customer engagement[13].

Opportunities

The interview responses showed that there were few opportunities from the current data collection system. Customers are more likely to openly express their opinions online than on a survey[13]. As more of the USPTO's customers shift to online communication, the USPTO's data collection system must follow. Most of the information from the current data collection system is very narrow. Thus, there was a need to explore a larger variety of feedback from the customer population. The assistance provided by passive data collection can improve the USPTO's understanding of the customers' perceptions.

Threats

The largest threat affecting the USPTO's current data collection system was that only USPTO employees rather than customers are frequently surveyed . This means that there is a large amount of customer feedback that is not being considered under this system[13]. One participant said they believed that stronger inclusion of minorities during outreach efforts could improve the quality of the feedback that is received by the USPTO. Beyond reaching more customers, this shows that receiving feedback largely from employees can lead to misleading conclusions about customers.

Selecting Data Sources

The team aimed to aid the USPTO in their efforts to better understand the perceptions of its customers during the patent application process. Research on types of tools and strategies for finding customer perception information on the internet was conducted. Although these tools are valuable, they cannot locate the desired feedback unless they are utilized in appropriate places. Therefore, it is important for the USPTO to understand not only how to collect information on customer perceptions, but also where to find that information. Outlined in this section are examples of locations that either reference the USPTO or which could yield helpful information if properly searched.

The team started by interviewing individuals who had patent experience but were not connected to the USPTO. One website that presented itself prominently was a patent blog named Patently-O. A major benefit to Patently-O is its popularity. The USPTO will be able to find more feedback opportunities on sites where there are enough posts. Patently-O has a page dedicated to journals that go in depth on a variety of patent topics, as well as sources for continued research. The site also has a drop-down menu with well over 100 categories that each contain many entries.

After the interviews, the team searched the internet for some other patent blog sites. One blog found was IPWatchdog. Like Patently-O, IPWatchdog also has a large library of patent experiences. IPWatchdog has an entire section on its site dedicated to posts involving the USPTO. IPWatchdog does not have as many searchable categories as Patently-O, but the categories still have plenty of content, and many have links to other topics.

In addition to professional blogs, the team also found Facebook groups that discuss patents. Patent Pals is one such group. Patent Pals has community standards requiring all discussions to be about patent law or other patent matters. An advantage to a community structured in this way is that data collection tools would not need to be as careful sifting through off-topic responses.

Using the previously discussed data collection tools, the USPTO will be able collect a wider range of data to better understand customer perceptions. However, it is also important for the USPTO to understand that it needs to be intentional with where it uses these tools. These websites are valuable places to conduct data collection due to the ratio of relevant comments to irrelevant comments.

One example of a data source that can provide actionable information to the USPTO is an article found by the team (without assistance from a web crawler) on IPWatchdog shown in Figure 14. The article discusses common issues surrounding 101 rejections. The full version can be found online. Highlighted in red boxes are key phrases and words that a web scraper could pick up. The words picked up from this document can become part of a larger collection of words from other documents. The accumulated text can then be used by an analysis method to uncover larger topics. If many similar articles were found, then topics such as difficulty explaining abstract ideas in

ineligible subject matter. The patent practitioner could get into a back-and-forth argument with the examiner about whether or not claim limitations are significantly more than the abstract idea, wasting multiple cycles of examination because examiners can be very difficult to convince of this. One promising approach is to argue that the claims are directed to a specific technological solution to a specific technological problem, as has been successful in the courts. But, even this may not be convincing, if argued in the abstract, because, after all, we are dealing with abstract ideas to begin with, and it is all too easy for an examiner to dismiss an abstract argument as "not convincing".

Figure 14: Example of Actionable Information

An example of actionable information the USPTO can look for as keywords in its search

a patent application or a waste of time by having abstract definitions in the patent law. This kind of information provides the USPTO more input on customer sentiments and with more data they can properly assess their level of customer satisfaction.

Organizing Attributes of each Technique

A decision matrix was used to evaluate the sentiment analysis strategies explored in the literature review. The literature review research revealed twelve predominant criteria which are laid out in Figure 15. The criteria chosen highlighted the most important factors in developing a sentiment analysis system. The items in brown represent software availability, dark blue is data preprocessing, and light blue is performance. Strategies were then ranked 1-6 for each criterion within a category, with 1 being the most applicable practice to the USPTO's needs and 6 being the least applicable. Ties were allowed within categories. Lastly, the total values for each column were added. The analysis strategies ranked, from lowest to highest in terms of the chosen criteria were LDA, LSA, SVM, ALA, NB, LA.

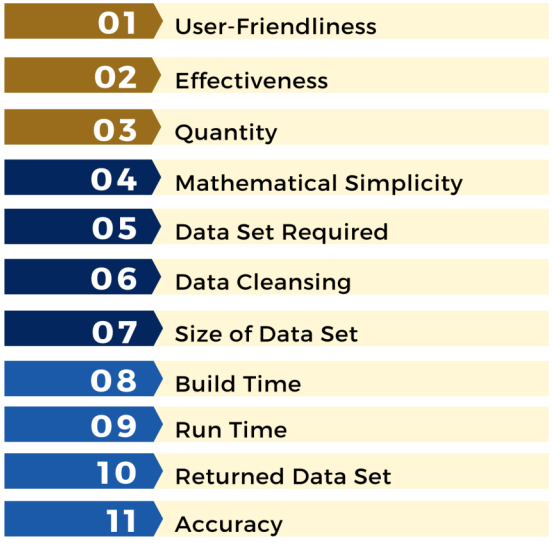


Figure 15: Decision Matrix Criteria

A list of criteria considered when completing the decision matrix found in the supplementary materials.

Recommendations

The team recommended to the USPTO a high-level strategy for collecting and analyzing passive data with the intent of discovering trends in their customers' perceptions. Specifically, the USPTO should:

1. Employ a passive data collection and analysis plan;
2. **Increase customer engagement through social media.**

With this recommendation is a literature review of best practices, a decision matrix of techniques and practices, a SWOT analysis of the current system, and a flowchart to visualize the data sources and techniques. The following section explains the team's rationale for this recommendation and these deliverables.

Initially, the team understood the data collection techniques to include data mining, text mining, web crawling, web scraping and web API's. However, through further research the team realized that the terms data mining and text mining are often used interchangeably and do not represent a concrete technique and were therefore removed. Web crawling and web scraping appeared to be highly related. Some software that claimed to be a web crawler would have the same functionality as another software that claimed to be a web scraper. The team decided to define the terms based on what was different between them. Web crawlers generally collect URLs of websites that contain keywords or phrases, similar to a google search. Web scrapers gather text from specific websites and can collect unordered words or full sentences. Web APIs are essentially web scrapers that are premade by the owner of the website to deliver

text information about that website. Once the team understood how these data collection techniques were related, the team created an order of operations.

Employing a Passive Data Collection and Analysis Plan

This order of operations is:

1. **Web crawl to gather a list of URLs;**
2. **Web scrape each URL for relevant text;**
3. **Place the gathered text into a database;**
4. **Using a data analysis technique to discover patterns in the data.**

When following this order of operations, there are several options to extract the data and ways to interpret it. A decision matrix and flowchart of the options were created to provide a simple and encompassing reference. These were required to fit the needs of varying conclusions drawn from different methods. Testing will be needed in the future to perfect the system. Each individual approach has strengths and weaknesses and is used to perform passive data collection with different foci in mind. For the USPTO to find new sources of information from which to collect data, **the team recommends locating the sources using a web crawler,** and then **performing the collection using a web scraper.** In many cases, a web API can be used to help reduce the complexity of the process used to find specific information. After the data are collected, the use of an analysis technique can give meaning to the information located. If the USPTO wants to understand the sentiment of its customers' posts, then **the team recommends Latent Semantic Analysis and Aspect Level Analysis.** If the USPTO wants to categorize found data into topics, then **the team recommends using Latent Dirichlet Allocation, Link Analysis, Support Vector Machines, or Naive Bayes Classifiers.**

To assist the USPTO in the collection of passive data, the team needed to understand how the current system worked. To understand this system, the team conducted a SWOT analysis. This SWOT analysis was completed by interviewing employees at the USPTO whose identities were rendered anonymous . The results from the SWOT analysis provided the USPTO insights into areas where they can improve and areas where they already succeed. This information will guide the creation of keywords that can be used to search with a web crawler.

When attempting to understand customer perceptions through website text, there is always the issue of overcoming what the motives were with the statement. Sarcasm and dishonesty will be unavoidable issues since there is nothing in text to symbolize such usage or intent. Extreme and frequent negativity will be more difficult to avoid but not impossible. For example, mechanisms exist within the software industry, such as machine learning, to recognize when the same group of individuals posts false or highly negative reviews, often called "trolling." Looking forward, **more research needs to be conducted in using machine learning to recognize topics and for finding new websites to investigate**. Machine learning can also be used to better understand the intentions of the customers.

Increase Customer Engagement Through Social Media

A secondary recommendation was for the USPTO to actively participate in online environments where it can engage with its customers. This will further assist in gathering customer feedback. A lack of social media usage became clear after interviews with the USPTO employees. Multiple interviewees mentioned that the USPTO does not reach out to many non-frequent filing customers. Besides the USPTO's website occasionally prompting for feedback, their Facebook, Twitter, YouTube and Instagram accounts displayed little interactions with customers. Improving the USPTO's online presence will allow for greater opportunities to gather passive data. The team's research briefly involved social media, however more exploration is needed into the nuance of improving the USPTO's presence online.

Implementation Plan

A flowchart was created and placed into the supplemental files. The flowchart details the primary decisions involved when building a passive data collection system. It lays out how to decide what collection strategy is required based on the experience of whoever is creating the system. It shows what factors to consider when manually finding data sources, and based on the results, how to preprocess the data. The flowchart shows which type of text classifier to use based on whether the user wants sentiment or topics.

Acknowledgements

Our team would like to thank our wonderful project sponsor, The United States Patent and Trademark Office, for their guidance during the course of this project. We would especially like to thank Martin Rater, Daniel Sullivan, David Fitzpatrick, Alix Eggerding, and Chelsea D'Angona.

We would also like to express our gratitude for everyone who participated in our interviews. Your responses were paramount to the success of this project.

Lastly, we would like to extend our thanks to our project advisors, Holly Ault and James Hanlan, for their unwavering support throughout the whole IQP experience.

Endnotes

1. *U.S. Patent Statistics Summary Table, Calendar Years 1963 to 2019, 05/2020 update*. (n.d.). Retrieved December 4, 2020, from https://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm
2. M. Rater, 2020, Passive Data Collection to Improve Service at the USPTO, USPTO
3. Patents External Quality Survey FY20Q4 Key Findings. United States Patent and Trademark Office, Sept. 2020.
4. D'Angona, C. (2019). *4-Customer Experience*. 19.
5. IAPP. "Passive Data Collection." 2020. Web. 14 Oct. 2020. IAPP. "Passive Data Collection." 2020. Web. 14 Oct. 2020. https://iapp.org/resources/article/passive-data-collection/#:~:text=Data%20collection%20in%20which%20information,other%20types%20of%20identification%20mechanisms.
6. Landers, R.N., Brusso, R.C., Cavanaugh, K.J., & Collmus, A.B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the internet for use in psychological research. *Psychological Methods*, *21*(4), 475–492. Scopus. https://doi.org/10.1037/met0000081
7. What is Intellectual Property (IP)? (2020). Retrieved from https://www.wipo.int/about-ip/en/
8. Information concerning patents. (2020, June 01). Retrieved September 10, 2020, from https://www.uspto.gov/patents-getting-started/general-information-concerning-patents

9. *Patent process overview*. (n.d.). [Text]. Retrieved December 4, 2020, from https://www.uspto.gov/patents-getting-started/patent-process-overview
10. Dijkmans, C., Kerkhof, P., & Beukeboom, C. J. (2015). A stage to engage: Social media use and corporate reputation. *Tourism Management*, *47*, 58–67. https://doi.org/10.1016/j.tourman.2014.09.005
11. Satisfaction Vs. Sentiment. (2014, December 15). *The Story of Telling*. https://thestoryoftelling.com/satisfaction-vs-sentiment/
12. Elkhani, N., & Bakri, A. B. (n.d.). *REVIEW ON "EXPECTANCY DISCONFIRMATION THEORY" (EDT) MODEL IN B2C E-COMMERCE*. 13.
13. USPTO Employee, in an interview with the group, November 2020.
14. Verhoef, P. C., Lemon, K. N., Parasuraman, A., Roggeveen, A., Tsiros, M., & Schlesinger, L. A. (2009). Customer Experience Creation: Determinants, Dynamics and Management Strategies. *Journal of Retailing*, *85*(1), 31–41. https://doi.org/10.1016/j.jretai.2008.11.001
15. Mozilla. (2019, September 29). An overview of HTTP. Retrieved October 09, 2020, from https://developer.mozilla.org/en-US/docs/Web/HTTP/Overview
16. What is Web Scraping and How Does Web Crawling Work? (n.d.). *Scrapinghub*. Retrieved November 16, 2020, from https://www.scrapinghub.com/what-is-web-scraping/
17. Lyu, F., & Choi, J. (2020). The Forecasting Sales Volume and Satisfaction of Organic Products through Text Mining on Web Customer Reviews. *Sustainability*, *12*(11), 4383. https://doi.org/10.3390/su12114383
18. *Machine Learning: What it is and why it matters*. (n.d.). Retrieved November 6, 2020, from https://www.sas.com/en_us/insights/analytics/machine-learning.html
19. Matta, P., Sharma, N., Sharma, D., Pant, B., & Sharma, S. (2020). Web scraping: Applications and scraping tools. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(5), 8202–8206. Scopus. https://doi.org/10.30534/ijatcse/2020/185952020
20. Xu, X. (2020). Examining an asymmetric effect between online customer reviews emphasis and overall satisfaction determinants. *Journal of Business Research*, *106*, 196–210. https://doi.org/10.1016/j.jbusres.2018.07.022
21. *What are extensions?* (n.d.). MDN Web Docs. Retrieved November 23, 2020, from https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/WebExtensions/What_are_WebExtensions
22. Grant, M. (n.d.). *How SWOT (Strength, Weakness, Opportunity, and Threat) Analysis Works*. Investopedia. Retrieved December 4, 2020, from https://www.investopedia.com/terms/s/swot.asp
23. *AnyPicker Tutorials: Visual Web Scraper | Web Crawler | Web Data Extractor | Web Data Visualization*. (n.d.). Retrieved November 24, 2020, from https://anypicker.ryangstudio.com/tutorials
24. Google. *Chrome Web Store*. Google. https://chrome.google.com/webstore/search/scraper?_category=extensions.
25. *PyPI · The Python Package Index*. (2020). PyPI. Retrieved December 4, 2020, from https://pypi.org/
26. *GitHub: Where the world builds software*. (2020). GitHub. Retrieved December 4, 2020, from https://github.com/
27. Selectors—Scrapy 2.4.1 documentation. (2020). Retrieved November 24, 2020, from https://docs.scrapy.org/en/latest/topics/selectors.html#topics-selectors
28. What is Web Scraping and How Does Web Crawling Work? (n.d.). *Scrapinghub*. Retrieved November 16, 2020, from https://www.scrapinghub.com/what-is-web-scrapin/
29. Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, *59*, 467–483. https://doi.org/10.1016/j.tourman.2016.09.009
30. *What is an API? In English, please.* (2019, December 19). FreeCodeCamp.Org. https://www.freecodecamp.org/news/what-is-an-api-in-english-please-b880a3214a82/
31. *Introduction to web APIs*. (n.d.). MDN Web Docs. Retrieved November 24, 2020, from https://developer.mozilla.org/en-US/docs/Learn/JavaScript/Client-side_web_APIs/Introduction
32. Trappey, A. J. C., Trappey, C. V., Fan, C. Y., & Lee, I. J. Y. (2017). Mining the Customer's Voice and Patent Data for Strategic Product Quality Function Deployment. In C. H. Chen, A. C. Trappey, M. Peruzzini, J. Stjepandic, & N. Wognum (Eds.), *Transdisciplinary Engineering: A Paradigm Shift* (Vol. 5, pp. 985–992). Ios Press.
33. Landauer, T. K., Foltz, P. W., & Laham, D (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259-284. https://doi.org/10.1080/01638539809545028

34. *Fit LSA model—MATLAB fitlsa*. (n.d.). Retrieved November 17, 2020, from https://www.mathworks.com/help/textanalytics/ref/fitlsa.html
35. Lettier. (2019, May 31). *Your Guide to Latent Dirichlet Allocation*. Medium. https://medium.com/@lettier/how-does-lda-work-ill-explain-using-emoji-108abf40fa7d
36. Blei, D. M. (2003.). *Latent Dirichlet Allocation*. 30.
37. Bi, J.-W., Liu, Y., Fan, Z.-P., & Cambria, E. (2019). Modelling customer satisfaction from online reviews using ensemble neural network and effect-based Kano model. *International Journal of Production Research*, *57*(22), 7068–7088. https://doi.org/10.1080/00207543.2019.1574989
38. *Link analysis: The lynchpin to better investigations*. Retrieved November 12, 2020, from https://www.visallo.com/blog/link-analysis-better-investigations/
39. Disney, . (2020, January 30). *Link analysis for fraud detection: A step-by-step example*. Cambridge Intelligence. https://cambridge-intelligence.com/link-analysis-fraud-detection/
40. Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), *Machine Learning: ECML-98* (Vol. 1398, pp. 137–142). Springer Berlin Heidelberg. https://doi.org/10.1007/BFb0026683
41. Grljevic, O., & Bosnjak, Z. (2018). Sentiment Analysis of Customer Data. *Strategic Management*, *23*(3), 38–49. https://doi.org/10.5937/StraMan1803038G
42. Guan, F., Shi, J., Cui, W., Hong, D., & Wu, J. (2019). A method for false alarm recognition considering threshold. *2019 International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC)*, 1043–1049. https://doi.org/10.1109/SDPC.2019.00199
43. *An Introduction to Support Vector Machines (SVM)*. (2017, June 22). MonkeyLearn Blog. https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/
44. *Module: Tf | TensorFlow Core v2.3.0*. (n.d.). Retrieved November 24, 2020, from https://www.tensorflow.org/api_docs/python/tf
45. Torres-Barrán, A., Alaíz, C. M., & Dorronsoro, J. R. (2021). Faster SVM training via conjugate SMO. *Pattern Recognition*, *111*, 107644. https://doi.org/10.1016/j.patcog.2020.107644
46 .Gandhi, R. (2018, May 17). *Naive Bayes Classifier*. Medium. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c
47. Mushtaq, M. S., Augustin, B., & Mellouk, A. (2012). *Empirical study based on machine learning approach to assess the QoS/QoE correlation* (p. 7). https://doi.org/10.1109/NOC.2012.6249939
48. *10 Most Popular Machine Learning Software Tools in 2020 (updated) | by Sophia Martin | Towards Data Science*. (n.d.). Retrieved November 15, 2020, from https://towardsdatascience.com/10-most-popular-machine-learning-software-tools-in-2019-678b80643ceb
49. *Naive Bayes—RDD-based API - Spark 3.0.1 Documentation*. (n.d.). Retrieved November 22, 2020, from https://spark.apache.org/docs/latest/mllib-naive-bayes.html
50. *Expectation confirmation theory—IS Theory*. (n.d.). Retrieved November 10, 2020, from https://is.theorizeit.org/wiki/Expectation_confirmation_theory
51. Tsao, W.-Y. (2013). Application of Expectation Confirmation Theory to Consumers' Impulsive Purchase Behavior for Products Promoted by Showgirls in Exhibits. *Journal of Promotion Management*, *19*(3), 283–298. https://doi.org/10.1080/10496491.2013.770811
52. Machado, M. J. C. V. (2019). Determinants of customer satisfaction: Empirical study in hotels. *International Journal of Applied Management Science*, *11*(2), 91–112. Scopus. https://doi.org/10.1504/IJAMS.2019.098823
53. Peacock, S. E. (2014). How web tracking changes user agency in the age of Big Data: The used user. *Big Data & Society*, *1*(2), 2053951714564228. https://doi.org/10.1177/2053951714564228
54. Taylor, I. (2017). Data collection, counterterrorism and the right to privacy. *Politics, Philosophy & Economics*, *16*(3), 326–346. https://doi.org/10.1177/1470594X17715249
55. Nunan, D., & Di Domenico, M. (2013). Market Research and the Ethics of Big Data. *International Journal of Market Research*, *55*(4), 505–520. https://doi.org/10.2501/IJMR-2013-015
56. Younes, A. S. (2019). Passive violation of consumers' privacy rights on the internet in the age of emerging data capital. *Journal of Content, Community and Communication*, *10*(5), 134–150. Scopus. https://doi.org/10.31620/JCCC.12.19/14