NER of Citations and Fine-Grained Classification of Responsa

Presented to the S. Daniel Abraham Honors Program

in Partial Fulfillment of the

Requirements for Completion of the Program

Stern College for Women

Yeshiva University

August 23, 2019

Adina Cohen

Mentor: Professor Joshua Waxman, Computer Science

Introduction

Jewish law (*halacha*) is rich and complex. It is a series of laws and practices that stem from the two parts of the Torah: the written text and the oral text. The written text is the Bible, the oral text is more amorphous but was compiled after the destruction of the Second Temple and was redacted in the third century CE in the text that is now called the Mishna. Following the compilation of the Mishna, the rabbis of the period, *Amoraim*, shifting from deriving laws directly from the written text. Instead they focused on expounding the texts of the rabbis of the time of the Mishna who were known as the *Tanaim*. The *Amoraim* explored the Mishna as well as statements of the sages that had not been written down called *Breitot* and those had been written down in a different work called the *Tosefta*. The teachings of the *Amoraim* and their discussions of the Mishna was eventually recorded as a commentary to the Mishna. This combination of Mishna with its Amoraic commentary is known as the Talmud. Just as with the compilation of the Mishna, the redaction of the Talmud represents a shift in the approach towards Jewish law. The shift following the redaction of the Talmud is seen as much greater than the one following the Mishna as it brought about the system of learning that still prevails to this day – a system that is focused on the study of the Talmud and trying to understand what the rabbis in the Talmud meant when they said one statement or another. It is during the initial period following the redaction of the *Talmud*, the period of the *Gaonim*, that the first halachic responsa appear.

Responsa (in Hebrew *She'elot v'Tshuvot* – Questions and Answers) are questions in Jewish law that are posed to the rabbinic authorities of the time. The rabbi then proceeds to trace through all the sources and ultimately come up with an answer to the question asked. As Jewish history continues and the Jews spread across the globe, the number of sources discussed

increases. This is because while the discussion starts at the Talmud, it extends to discuss the sources that comment on the Talmud as well as traditions that were passed down in various parts of the world (mainly Eastern and Western Europe).

Responsa give a fascinating lens into the issues that specific communities were facing throughout Jewish history. Questions can be as tame as a question in the laws of the Sabbath to questions fraught with emotion such as the ones posed to rabbis from within the horrors of the Holocaust. Studying responsa on a large scale, therefore, provides rich insight into Jewish history.

HaMapah, headed by Ellie Fischer and Moshe Schorr, is a project that aims to harness the power of technology to analyze responsa. The name "HaMapah" is a play on words. In modern Hebrew, "HaMapah" means "the map" which is an apt name for a project that explores Jewish communities across the globe. However, "HaMapah" is also the name of the gloss to one of the most influential halachic works of all time, the *Shulchan Aruch*. In the context of the Shulchan Aruch, which literally means "set table", HaMapah means "tablecloth" and was meant by Rabbi Moshe Isserles , the Rama, to fill in gaps where the religious practice differed in the communities in which the Rama lived from that which was written in the Shulchan Aruch.[1] The HaMapah project specifically looks at questions posed from one great rabbi to the next in roughly the nineteenth century in Europe (Schorr, 2018). Fischer and Schorr focus on understanding the power and reach of specific rabbis. They initially analyzed responsa to see where questions were posed from and drew heat maps of the results. This allowed them to see how far the reach of a given rabbi extended as compared to other rabbis of their time. For example, HaMapah analyzed the responsa of Rabbi Shalom Mordechai Schwadron of Berezhany who lived in Galicia from

[1] This pun was pointed out to me by my mentor, Professor Waxman

1835 – 1911 and is commonly known as the Maharsham. While the vast majority of questions asked were from Galicia, it is fascinating to see that the Maharsham even received questions from places miles away such as France, England and Istanbul (Schorr, 2018). This analysis, therefore, proves just how much of an influence the Maharsham had during his generation (Figure 1).

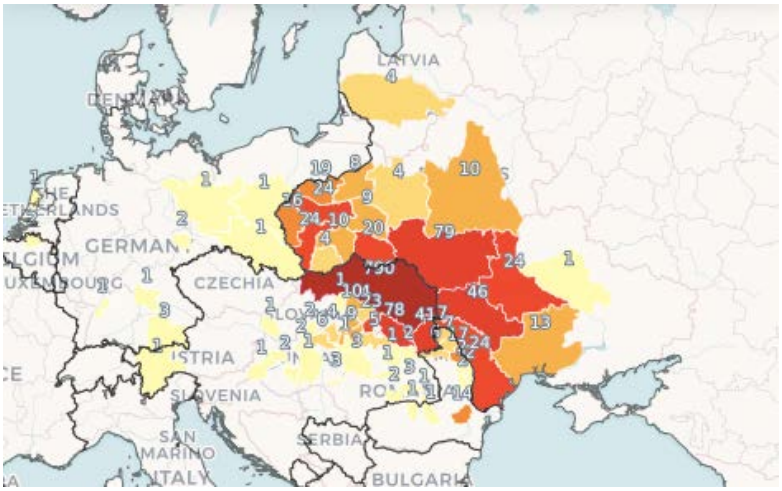

Figure 1: Heat Map of the Reach of the Maharsham (Schorr, 2018)

HaMapah, however, has realized that looking simply at the number of questions posed to a specific rabbi does not accurately reflect his authority. In a blog post about this issue, Schorr wrote,

> But are all responsa created equal? If a gabbai asks a rav whether the congregation should skip tachanun[, a prayer that represents mourning,] on Erev Tu BiShvat[, the eve of the new year for the trees], and the rav sits down and writes a lengthy treatise in response, does it really tell us anything about his authority? There are certain types of questions that demonstrate real influence. If people carry on Shabbat based on an eruv approved by a particular rabbi, over and against competitors, it indicates authority. The higher the stakes, and the more lives the question affects, the more important the responsum, and the more authority demonstrated by the responding rabbi. (Schorr, 2018)

In order to truly understand the authority of a rabbi, the content of the questions he is asked is extremely important. By looking at the content, it can be determined whether the question posed

is a “simple” or “weighty” question. Since it is impractical to look at every single question and decide how “weighty” the question was, it is possible to pinpoint the topics in which weighty questions are generally found.

However, responsa, especially those found online, are not neatly organized by topic. Thus, in order to take the analysis of responsa a step further and be able to garner information about the actual content of the questions being posed to rabbis, there has to be a way to split responsa into meaningful topics.

The goal of this project was to write a program that would successfully split responsa into meaningful topics using Named Entity Recognition (NER). Before delving into the coding approach to the question at hand, the goal had to be better defined. What is considered a meaningful topic?

Splitting the Responsa

If there was a shift in the way that rabbis approached Jewish law after the Talmud, a second shift occurred in the mid 16th century. In the 13th century, a Sephardic rabbi by the name of Rabbi Yaakov ben Asher (the Tur) wrote a halachic code called the *Arbah Turim*. The work took all of Jewish law and split it into four sections: *Orach Chayim, Yoreh De’ah, Even HaEzer* and *Choshen Mishpat*. *Orach Chayim* deals with laws of daily life, *Yoreh De’ah* with dietary restrictions, *Even Ha’Ezer* with marriage and divorce and *Choshen Mishpat* with damages and finance (Segal E., n.d). A few hundred years later, Rabbi Yosef Karo wrote a commentary on the *Arbah Turim* called the *Beit Yosef*. He then took the “bottom lines” of his commentary and compiled it into a halachic code that he called the *Shulchan Aruch*. This code represented another shift in Jewish law. Following the publishing of the *Shulchan Aruch* which was accompanied by a gloss from an Ashkenazi rabbi, Rabbi Moshe Isserles, the code was almost universally

accepted.[2] From the 16th century onwards, all questions of Jewish law are answered through referencing the *Shulchan Aruch* in one form or another. As such, the sections used by the *Tur* and subsequently by the *Shulchan Aruch* are commonly used throughout halachic works. Therefore, it would be logical to assume that any meaningful categorization of responsum should follow the sections of the *Shulchan Aruch* in one form or another.[3]

Beyond the four sections of the *Shulchan Aruch*, each section is further split into sections called *simanim*. These *simanim* were split into groups by the *Tur* (and followed in the *Shulchan Aruch*) and labeled by topic. For example, the first seven *simanim* of the *Orach Chayim* section in the *Tur* and *Shulchan Aruch* discuss the laws of waking up in the morning. These topics, already defined by the *Tur* and *Shulchan Aruch*, are the topics that are most meaningful when it comes to categorizing responsum.

Although there are four sections of the *Shulchan Aruch*, the project focused specifically on the *Orach Chayim* section which discusses daily life, prayer, Shabbat and holidays. Since the method of analyzing the text will be the same for each of the four sections, this project can easily be adapted to fit one of the other sections of the *Shulchan Aruch*.

Orach Chayim contains 697 *simanin* which are split into forty meaningful topics. It is into these topics that the project aims to fit the responsa analyzed.

Methods

There are two different ways in which I aimed to classify the responsa. Both approaches employ the use of a text other than the responsum at hand. Originally, I chose the *Shulchan*

[2] The biggest exception to this during the time of the Shulchan Aruch was the Maharshal – Rabbi Shlomo Luria. His aversion towards the codification of Jewish law is further explained in an article by Rabbi Shlomo Brody: http://text.rcarabbis.org/against-the-shulchan-aruch-the-critique-of-the-maharshal-by-shlomo-brody/.

[3] Many Jewish Rabbis are referred to colloquially by their most famous work. Thus Rabbi Yosef Karo becomes the Shulchan Aruch and Rabbi Yaakov ben Asher becomes the Tur.

Aruch as the text of choice since as mentioned above the *Shulchan Aruch* is the text on which most halachic questions are decided. However, the *Shulchan Aruch* is a compilation of bottom lines. Therefore, the amount of words spent on a given topic is extremely small when compared the number of words in a responsum. This is because there are two major ways in which responsa are different from the *Shulchan Aruch*. The first is that the *Shulchan Aruch* brings down the bottom line whereas a responsum goes through the entire discussion from the Torah and Talmud down to the rabbis who have discussed the topic throughout the generations. The second is that the question asked in responsa are usually questions of complexity that cannot be answered by simply opening up the *Shulchan Aruch* and looking up a particular law. Responsa must reach back to earlier sources and are therefore lengthier. Thus, instead of using the *Shulchan Aruch*, I used its precursor, the *Beit Yosef*. The *Beit Yosef* is helpful because similar to responsa, it goes through the discussion that leads to final law.

I used the *Beit Yosef* in two ways and both approaches will be discussed at length. Below is an overview of the approaches.

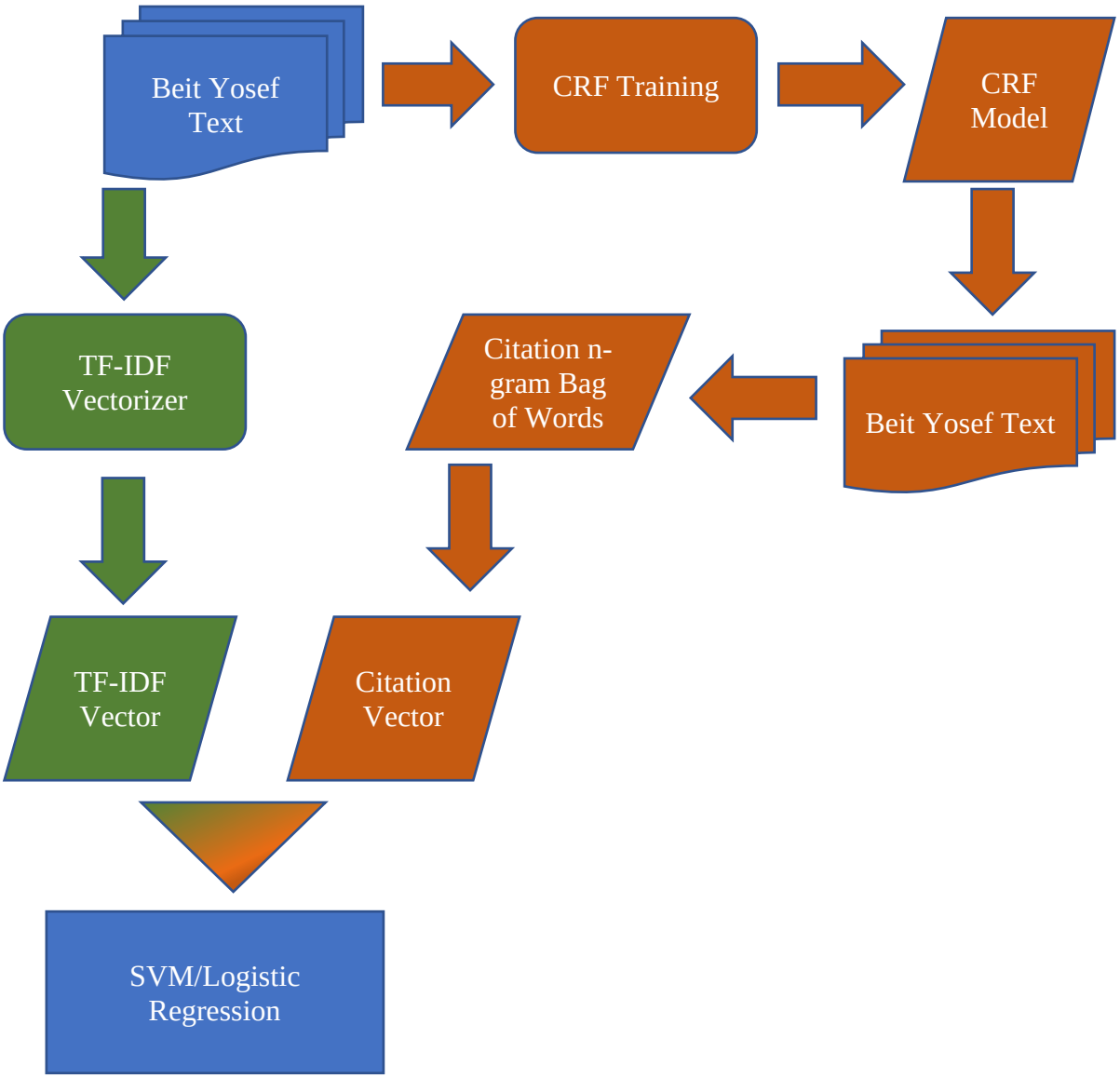

Figure 2: Overview of Approaches

The first approach (the green boxes in Figure 2) employed the text of the *Beit Yosef* itself and built an SVM of n-grams of the text and then used that SVM to make guesses about the responsa. In this approach, the text of the *Beit Yosef* was fed into a TF-IDF vectorizer and then those vectors were fed into the SVM. This approach made use of the Scikit-Learn module from Python will be referred to as the straight-text approach.

The second approach (the orange boxes in Figure 2) makes use of the citations found in the *Beit Yosef*. Throughout his commentary on the *Tur*, the *Beit Yosef* generally cites passages from Talmud and opinions of commentators. These citations are likely to be found in responsa as the rabbi writing a response to a question will also trace through the sources from the Talmud until his present day. Citations were therefore extracted from the *Beit Yosef*. This was done by building a conditional random field (CRF) model. The model was then used to build a bag of words filled with n-grams of citations. The bag of words was then added to the vectors produced in the straight-text approach and added to the analysis of the SVM. This approach made use of the pycrfsuite module from Python will be referred to as the citation approach.

I. The Straight-Text Approach

When looking at a sample of text and trying to classify it, there are many different methods that can be used to go about this process. The simplest method involves using a "bag of words". When classifying, while it may seem instinctive to look at a full text, in truth, many of the words in the text are useless and will not help with classification. For example, in the sentence, "The cat ran around the room looking for the mouse.", the words "the" and "for" will not help in figuring out the meaning of the sentence. The words, "mouse", "cat", "room", and "looking", however, are extremely meaningful. These words are extracted from the sentence and described as features of the text. The features are what ultimately make up the bag of words. The bag of words can then be used in different ways in order to classify text. We will walk through an example of a teshuva in order to model two different ways of using a bag of words.

שאלה אם שכח ע"ה בסעודת פורים אי צריך לחזור או לא

תשובה יראה דצריך לחזור וכן נמצא במרדכי בכתב יד שהגיה גדול בדורו מהר"ר יעקב פולק ז"ל שמתחילה כתב בחנוכה ופורים אין מחזירין אותו בעל הניסים ולבסוף כתב שבת וי"ט ופורים דלא סגי דלא אכיל מחזירין:

(*Teshuvot Maharshal Siman* 48)

The example text above is a question about whether a person should repeat *Birkat HaMazon*, Grace After Meals, on *Purim* if he/she forgot the special addition of *al hanisim*. Our goal is to try to see if this question is about *Purim*. For the purposes of this example we will be using a bag of words with five words and two bigrams – פורים, מגילה, ע״ה, לאביונים, סעודת, בסעודת פורים, מגילה פורים.

The first way to use the bag of words is a true/false bags of words. What this means is the program will simply check if each word in the bag of words is in the text and will mark it one if it is present and zero if it is not (Figure 3). In this case, the words פורים, ע״ה, and בסעודת פורים were found in the teshuva.

A second way in which to use a bag of words is a simple count bag of words. Using this bag of words, the numbers in the vector represent the amount of times each word appears in the text (Fig 3). The advantage of a word count vector is that it allows words that are repeated to hold more weight when being analyzed. In the vector below, while the majority of the numbers stay the same, the word פורים would receive more weight as it appears three times in the above text.

1	0	1	0	0	1	0
פורים	מגילה	ע״ה	לאביונים	סעודת	בסעודת פורים	מגילה פורים

3	0	1	0	0	1	0

Figure 3: True/False Bag of Words Vector and Simple Count Bag of Words Vector

Using these vectors, the program would then decide whether the text above should be classified as being about *Purim*. Both of these methods, however, are extremely simple and will not always produce the best results. The true/false bag of words is not accurate because it fails to reflect which words appear more frequently than others. The simple count bag of words is also not the most accurate because it does not account for words that may appear multiple times in the given document but also appear frequently in other texts of a different topic. Alternatively, it does not account for words that while they may appear infrequently in the given text, are words that are extremely significant and should be granted more weight. Thus, there is a third model that deals with these issues.

The third model involves looking at the words that are frequently found in one text and yet infrequently found in other documents. For example, if a person is trying to classify books by genre where some books are fantasy while others are realistic fiction, it would be helpful to find the significant words in each of the books. While it may seem obvious that words such as "magic", "sorcery" and "spells" indicate that a book belongs in the fantasy genre, the computer does not know that. Therefore, it needs to determine that these words are significant through processing the entire corpus. Word counts are done for a given text and then compared to the number of times it appears elsewhere in the corpus. If the word "magic" appears many times in a given text, it may be an indicator that it is a trigger word. This is proven to be true if is it also not found as frequently in other texts. Once trigger words are identified, it would be possible to classify a given text into the genre of fantasy by finding those trigger words in the text. This concept is known as "term frequency inverse document frequency" (TF-IDF) (Rajaraman & Ullman, 2011).

In order to be able to analyze texts using TF-IDF, vectors of each text are formed. The vectors are arrays in which each word in the corpus is represented through a TF-IDF score. The term frequency of a word is calculated by taking the word count of a term and dividing it by the maximum number of times any term appears in that document. This serves to normalize the word count.

$$TF_{ij} = \frac{f_{ij}}{max_k f_{kj}}$$

In the formula above the term frequency is equal to f_{ij} which is the frequency of the word *i* in the document *j* divided by the maximum occurrence of word *k* in document *j*.

This is then multiplied by the inverse document frequency – i.e the frequency with which a given term appears in the rest of the corpus where *N* is the number of documents in the corpus and n_i is the number of documents in which term *i* appears. (Rajaraman & Ullman, 2011).

$$IDF_i = \log_2(\frac{N}{n_i})$$

The TF and IDF scores are then multiplied together to product the TF-IDF score for a given word. (Rajaraman & Ullman, 2011). Each word in the vector represents a feature of that text. Since corpuses can contain hundreds of thousands of words, it is important to limit the number of words included in the analyses. In my analysis, I chose to create vectors with a max feature number of 50,000 which means that the 50,000 most significant n-grams were taken into account when classifying a given responsa.

Before moving on to the citation approach, there is another aspect of bag of words that needs to be discussed. In the initial example with the sentence about the cat, the bag of words contained single words only, i.e unigrams. In the teshuva example, however, the bag of words

also contained phrases of two words known as bigrams. Unigrams successfully capture words, but they do not capture as much meaning. Consider the following sentence: “She went to First County Bank to open a bank account.”. The words “First”, “County” and “Bank” all hold meaning in and of themselves, but if I were to analyze this sentence using only unigrams, I would be unable to pinpoint that there is a proper noun in the sentence that is the name of a bank. Using n-grams (phrases of more than one word) allows the capture of such elements. In this particular sentence, the unigrams would be each individual word. The bigrams would be “She went”, “went to”, “First County”, “County Bank”, “Bank to”, “to open”, “open a”, “a bank” and “bank account”. Just using the bigram without the trigrams adds a layer of understanding to sentence as now the phrase “bank account” can be analyzed together. By taking it one step further and analyzing the trigrams, the program would discover “First County Bank” and be able to tag that as the name of an organization. In both approaches in this project, we utilized unigrams and bigrams in order to increase the accuracy of classification.

II. Citation Approach

When answering complex halachic questions, rabbis look to earlier sources to help determine the correct course of action. Once the answer is reached and is being recorded, the entire logic is written down from start to finish in addition to the final *psak*, halachic ruling. This means that responsa contain many references to earlier sources such as passages from the Talmud as well as later rabbinic works. Utilizing these sources cited is another more complex way of determining the topic of the responsa.

The *Beit Yosef* also went through the sources from the Talmud until his bottom line, quoting and citing sources, therefore it made sense to use the *Beit Yosef* for this approach as well.

III. Annotating the Beit Yosef

Selections from the *Beit Yosef* were first manually annotated using a Java web annotation program called WebAnno. Fifteen random *simanim* were chosen to attempt to get as many variations of citations as possible. The data sample used was relatively small which can provide limitations in accurately extracting citations. At the same time, there were a total of 229 citations extracted from the fifteen *simanim* which is much more significant[4].

Each text was tagged with two different the labels. The first was the "Citation" label which pinpointed specific citations. The second was the "Citation_Introduction" label. This label was created to increase the accuracy of citations when they would later be extracted from the entire text of the *Beit Yosef*.

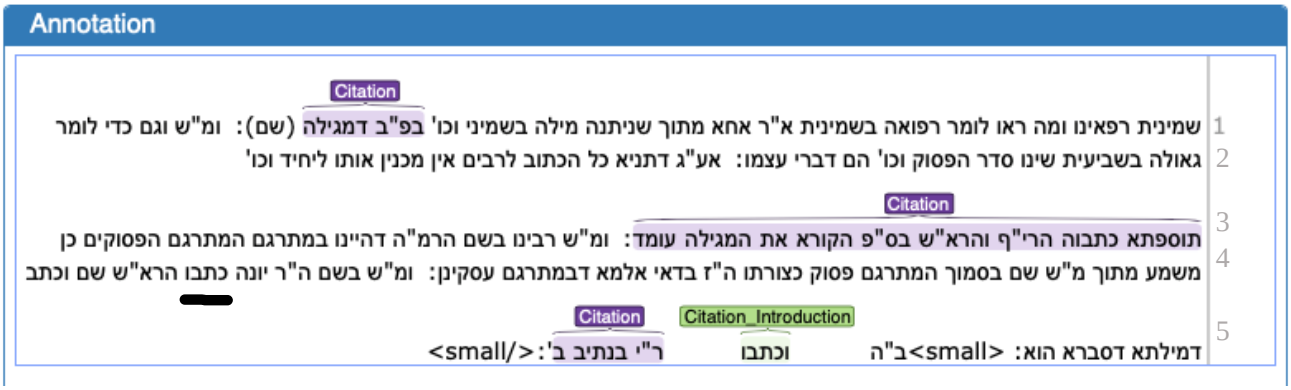

Figure 4: Sample of Annotated Beit Yosef Text

The sample text above is from the *Beit Yosef*'s comment on *Tur Orach Chaim* 116. There, the Tur discusses the eighth blessing of the *Shemonah Esrei*, the Eighteen Blessings, a part of the daily prayer. The eighth blessing, titled *refainu*, is a blessing in which the supplicant asks God to heal those who are sick. The Tur quotes a Talmudic passage that suggests that the

[4] It is also important to note that when scanning the list of words that ultimately make up the citations bag of words, the vast majority are valid citations. I did a quick count and only encountered a questionable citation after seeing 24 valid citations. Having a larger data size would potentially make the error margin even smaller. It would also allow for citations that may have been overlooked in the extraction process to be extracted.

reason that the blessing of *refeinu* is the eighth blessing is because circumcision is on the eighth day of a baby boy's life and healing is required after circumcision. The Tur, however, does not cite the source of this Talmudic passage. Rav Yosef Karo, therefore, tracks down the source and quotes it in line one.

The citation as annotated reads בפ״ב דמגילה which refers to the second chapter of the tractate of *Megillah*. The word שם follows in parenthesis. Generally speaking, שם in Jewish works are the equivalent of ibid. in English works; it is a reference back to the previous citation. Since this is the first line of the *siman*, referencing back to the previous *siman* will provide the full citation – (יז:) בפ״ב דמגילה. For the purposes of this project, I chose to ignore citations that appeared in the form of שם as incorporating them would be extremely intensive, however this aspect could be expanded upon in the future.

The second citation is a reference to a *Tosefta* brought down in the Rif and the Rosh, two commentators, at the end of fourth chapter of *Megillah*. In this case the citation does not explicitly mention the chapter number but rather the chapter name – "One Who Reads the *Megillah* (Scroll) Standing". The Rif is a commentator on the Talmud who takes the text of the Talmud and strips it down to the halachic bottom lines. Sometimes the Rif will add to the bottom lines and in this case, his additions are found in the form of a *Tosefta* that is not quoted in the Talmud. The Rosh is a commentator on the Rif. The *Beit Yosef* is therefore referencing the Rif and the Rosh's further comments on the Rif, in this particular part citation.

The third citation references the work of Rabbeinu Yona, a thirteenth century Rabbi. The citation introduction word "וכתבו" means "and he wrote it" which is a classic introductory word. It is interesting to note however, that this word could be used to introduce a reference without an explicit citation (see line 4 where the citation is the word שם). Thus, although it was initially

suggested that introductory words would aid in finding citations, it actually did not seem to make much of a difference at all. The reason could be because of this issue illustrated above.

Once the citations were tagged for those fifteen texts, the texts were exported so as to create a model to be able to extract citations from the rest of the *Beit Yosef*. The most important feature needed in an algorithm that can properly tackle this task is the ability to handle context. Many words have multiple meanings and connotations and it is only when looking at context that the true connotation of the word is understood. For example, the word *siman* (סימן) in Hebrew means sign.[5] *Siman*, however, is also the word used to indicate a section in a book; the *Beit Yosef* is split into sections called *simanim*. In order for the program to be able to tell whether the usage of the word סימן is a citation or not, the surrounding words must be examined. If the word is followed by a number then it would be a clear indication that the word is a citation.[6] Similarly, when tagging citations, looking at the context of previous words is helpful. If a word immediately follows a word that was tagged as a citation, it will be more likely to be tagged as a citation as well.

Given the requirement of context, certain models such as a Hidden Markov Model (HMM) or stochastic grammars could not be used. With a HMM or stochastic grammar, the model must determine all possible pairings of observations in order to find the highest probability of labels over a given observation sequence. (Lafferty et al., 2001) Take the word "Old" in the following sentence: "Sally ran away to Old Farmington". The word "old" is generally an adjective. Here, however, it is a proper noun. What the HMM would do is generate

[5] The proper use of the word *siman* through its translation into sign would be, "if there is no sign on the object of who the owner may be, the object is considered ownerless and can be claimed by anyone".
[6] In Jewish works numbers are expressed as letters. Each letter in the Hebrew alphabet has a value and letters are strung together to form numbers. For example, the current Hebrew year 5779 is expressed in Hebrew letters as התשע״ט. The ה here represents 5000 and then תשע״ט adds up to 779. ת = 400, ש = 300, ע = 70 and ט = 9.

all the possible tags of the sentence. One such set would have “old” tagged as an adjective and another would have it tagged as a proper noun. In order to determine which of the tag sets is most likely correct, the probabilities determined for each individual tag are multiplied together. The sentence with the highest probability is the chosen tag set. In a sentence with only six words such as the one above, generating all the tag possibilities is doable from a time complexity perspective. As the data size increases, however, number of possibilities increases astronomically. It is therefore impossible to actually make use of all the possibilities. What happens instead is that the decisions are made locally looking at the context of perhaps the surrounding few words but not the larger context. (Bird, Klein & Loper, 2009) Conditional random fields (CRF) provide a solution to the context problem demonstrated in HMMs.

Conditional random fields is a “framework for building probabilistic models to segment and label sequence data” (Lafferty, McCallum & Pereira, 2001). It is easiest to understand how a CRF works through a simple example. Imagine you went on a trip throughout Europe and now have a folder of hundreds of pictures. You would like to label all your photos with their locations using a computer algorithm. One way to classify the pictures would be to look at each image individually and try to classify based off of tagged pictures you scraped from the internet of all the cities you visited. Using such a method, pictures of the Eifel Tower, Buckingham Palace or perhaps even a busy street would be easy to label. Where the algorithm would struggle would be when it comes to the close-up pictures of you and your friends in the middle of a forest or standing against a wall. If, however, you looked at the previous picture and noted that it was a picture from Venice then it would make sense to give more weight to Venice when choosing how to label your image. This is because your pictures are in order of your travels and therefore

groups of pictures were taken in the same location. Looking at the sequence of the photos is a strategy that CRFs employ. (Chen, n.d.)

In this example, the CRF would make use of a transition matrix when deciding how to tag a given picture in addition to looking at other features of the image. The matrix contains the costs of the tag of an image changing from location X to location Y_i where X is the location of the previous image and Y_i is a location from Y which is the set of locations visited on your trip. Each image will then be not only tagged simply by looking at the features of the image itself, but also through referencing the costs of it being tagged a certain way based on the transition matrix. Essentially this transition matrix allows the conditional model to look at the broader picture more easily. (Boulton, 2018)

In the case of extracting citations, consider the case where the tagger comes across the word "*amud*" which can either mean chapter or pillar. It is possible that a rabbi is discussing laws related to building and uses the word "*amud*" to discuss the pillars of the building. However, if the word following "*amud*" is a word that represents a number or some of the previous words constitute the name of a book, the likelihood that "*amud*" is a citation increases. This ability to successfully use context in order to label data is one of the advantages of CRFs.

The CRF model built on the 15 tagged simanim predicted citations extremely well. It had a 97% precision score, 99% recall and 98% F1-score. While the precision for the citation_introduction label scored 100%, the recall and F1-score performed much worse with 16% and 27% respectively. The poor scores of the citation_introduction label, however, do not affect this project as the citation_introduction was only used to help discover citations. Moreover, the results of the model without the citation_introduction tag were approximately the

same (97% precision, 100% recall) and therefore as their presence did not negatively affect the results they were left in the model.

Once the CRF model was built, citations were extracted from the rest of the text of the *Beit Yosef*. The data was loaded into the program as long strings, so the *simanim* first needed to be converted into arrays of strings where each entry in the array was a word from the *siman*. In addition to creating a matrix with the entire text of the *Beit Yosef,* a similar matrix needed to be created with the features of the words in the text. The CFR tagger then tagged all of the words using the features matrix. Once all the words were tagged, the words that were specifically defined as citations were extracted into a bag of words. The most crucial part of the extraction of citations was to ensure that citations were not extracted word by word but instead as phrases. For example, if the *Beit Yosef* cited מגילה ד. we want to ensure that our bag of words reflects that, as having just מגילה (the tractate of Talmud) or ד. (the folio and page number) alone is not meaningful. In order to do this, I kept track of the previous tag and if the previous tag was also a citation, it was assumed that the two words were part of the same citation and recorded as such. This extraction resulted a list of 3002 citations. I then imported the list and created a bag of words containing the n-grams of the citations.

Using this bag of words, each text of the responsa was analyzed and a vector created. This vector was added to the end of the vectors created via the straight-text approach. Once the vectors were created for each text entry, the resulting vectors had to be analyzed. There are many different algorithms that have been written to classify texts. I employed the use of two different classifiers. The first is a stochastic gradient descent (SGD) implementation of a liner SVM and the second is a logistic regression classifier. The pros and cons of each classifier as well as which classifier performed better will be discussed in the results section.

IV. The Linear SVM Classifier

An SVM (Support Vector Machine) is a supervised learning tool used to classify items. It was originally built to handle two-group classification problems however it has been since modified to handle multi-group classification (Cortes & Vapnik, 1995). The scikit-learn documentation defines an SVM as a machine which

> constructs a hyper-plane or set of hyper-planes in a high or infinite dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyper-plane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. ("Support Vector Machines", n.d.)

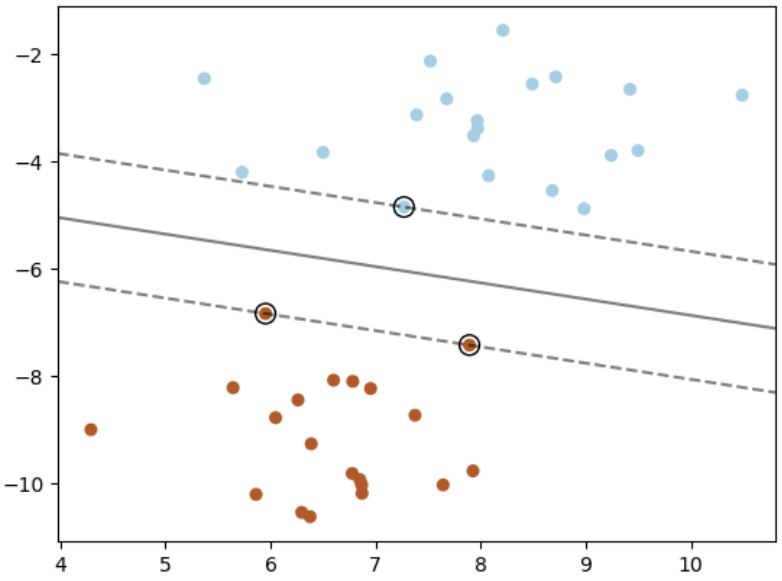

Figure 5: Simple Linear SVM (scikit-learn documentation 1.4.7)

This definition can be most easily explained through the image above. Imagine the plane without the line. All that exists are two groups of dots, the orange ones and the blue ones. The orange dots represent cities and the blue dots represent towns. These dots have already been manually classified. The goal, however, is to use this information to be able to take data about a random

place and be able to classify it as a city or town. An SVM takes the vectors that have been built for each item in the dataset and graphs them on a plane. Then the SVM attempts to separate the two sets of data by finding a hyperplane that successfully separates the set and provides the widest gap between the two sets. The wider the gap, the easier it becomes to classify new data. In other words, if the classifier can determine which features best distinguish cities from towns, it will be able to make the most educated guess when handed a new piece of data. The example above is a simplified one in which the data is only split in two. In this project that data was split into forty categories and therefore the vectors are divided with hyperplanes as opposed to simple lines.

V. Logistic Regression

Logistic regression is a probabilistic model of classification. Instead of choosing one correct class, the classification model will determine the likelihood of an item belonging to a given class. This means that it is possible to not only look at the classifier's top choice, but its successive guesses as well. This feature proved to be a big asset to the project.

VI. From the Beit Yosef to Responsa

Both models were trained and tested on the text of the *Beit Yosef*. The testing set size was ten percent of the overall dataset which means that the model was trained on 627 *simanim* and tested on the remaining seventy *simanim*. Since the results of the models were satisfactory when it came to classifying the text of the *Beit Yosef*, the model was saved and used to make classification predictions on a set of responsa.

The model was tested on twenty-four manually tagged responsa. Sixteen of the responsa were taken from the *Teshuvot Maharshal* and eight were taken from the *Noda B'Yehuda*. The *Noda B'Yehuda*, Rabbi Ezkiel Landau, lived in Poland in the 18th century. (Mindel, 2004). The

Teshuvot Maharshal were written by Rav Shlomo Luria and was a contemporary of Rav Yosef Karo. In fact, the Maharshal was against the widespread adoption of the *Shulchan Aruch*. (Brody, 2010). Classifying his responsa, therefore, emphasizes that the Maharshal may have been against the concept of the *Shulchan Aruch* but it was only because it removed a certain level of halachic independence from rabbis. The Maharshal, however, still drew from the same sources as the *Shulchan Aruch* when coming to conclusions of halachic nature and therefore the work of Rav Yosef Karo can be used to classify the work of the Maharshal. Both the Maharshal and the Noda B'Yeuda were used in order to test if the classifier works on responsa written by various authors from different time periods.

Results

There are two different sets of results for each classifier. The first is how well the classifier performed simply with using the text of the *Beit Yosef*, i.e how well it classified random *simanim* in the text. The second set of results is how well the classifier then classified the test responsa. The scores for each of the results is the mean accuracy score.

I. SGD Classifier

The SGD classifier received a score of 99.36% accuracy on the *Beit Yosef* training data. It received a score of 74.29% on the *Beit Yosef* testing data.

When run against the twenty-four responsa, the classifier achieved an accuracy of 62.5%. This means that it correctly classified fifteen of the responsa. The nine incorrectly classified responsa were classified as follows:

	Guess	**Answer**

1	Hannukah	*Purim*
2	*Eiruv*	*Shabbat*
3	Synagogue	*Chol HaMoed*
4	Hannukah	Passover
5	The Prayers of *Sukkot* etc.	*Sukkah*
6	The Prayers of *Sukkot* etc.	*Sukkah*
7	Hannukah	Grace After Meals
8	Waking in the Morning	*Shabbat*
9	Hannukah	*Purim*

Responsa numbers 2, 5 and 6 can be counted as being classified correctly as the laws of *Eiruv* are a subset of the laws of *Shabbat* and the Prayers of *Sukkot* can just as easily be defined as the laws of *Sukkot*. This brings the number of correctly classified responsa up to eighteen and increases the accuracy of the classifier.

In order to determine how far off base the classifier was for the other incorrectly classified responsa, the responsa must be looked at individually.

a. Responsum 1

אם המולד של ר"ח אדר היה בער"ח ואז כלה זמן קידוש לבנה בפורים באופן שליל י"ד שהיא ליל קריאת המגילה היא לילה אחרון שיכולים לקדש הלבנה ואם לא קדשוהו בו בלילה עבר הזמן. ולא נראית הלבנה כל אותן הימים ובליל י"ד באמצע קריאת המגילה נתפזרו העבים וזרחה הלבנה אם יש להפסיק באמצע קריאת המגילה לקדש הלבנה כי אולי אם ימתינו עד סוף הקריאה תתכסה בעבים:

(*Noda B'Yehuda - Orach Chaim Siman* 41)

There is a Jewish law to make a blessing over the new moon at the start of every month. While the blessing is generally made the first Saturday night after the month begins, if the moon cannot be seen, the blessing cannot be said. Therefore, a person has until the fifteen of the month to make the blessing on the moon. During the Hebrew month of *Adar*, the holiday of *Purim* takes place on the night of the 14th. On that night, the scroll of Esther is read. The question asked sets up a scenario where it is cloudy the whole first week and a half of the month of *Adar* and there is no opportunity to bless the moon. On the night of the fourteen, the last opportunity to make this blessing, the clouds suddenly part. However, this happens in the middle of the reading of the book of Esther. Should the congregation pause the reading in order to go outside and recite the blessing over the moon?

The classifier suggested that this question should be classified as question about Hannukah. While there are many similarities between the laws of Hannukah and *Purim*, perhaps explaining the misclassification, this question is clearly not about Hannukah and is classified entirely incorrectly.

b. Responsum 3

נשאלתי על דבר בנין הגג על ביתו של הקצין ר' ליב ק"פ שהוא לפ"ד נחוץ לו מאד ואם ישבתו הבעלי מלאכה כל ימי חוה"מ ישאר הבית זמן רב בלי מחסה ומסתור מזרם וממטר אח"כ בימות הגשמים כי אח"כ ילקחו כל הבעלי מלאכה לבנין המבצר ומי יודע עד כמה לא ישיג לשכור ב"מ, בכן רוצה להתיר לו לבנות בחוה"מ והוא בקיבולת:

(*Noda B'Yehuda - Orach Chaim Siman* 12)

This question discusses whether a person can build a drain on their roof during the middle days of the holidays (*Chol HaMoed*), a time when normally such work is forbidden. The reason given is that after the holiday ends the rainy season will start and the builders will be unable to build and the lack of proper drainage will cause issues for the house.

The classifier suggested that this question should be classified as the “Laws of Synagogue” but that is clearly an incorrect classification.

c. Responsum 4

בחטה שנמצאת בפסח בגרעמזליך שנעשים משומן ודבש ומים לפי שחכם אחד רצה להתיר בששים לפי דעת המג"א בסי' תמ"ז ס"ק ה' דחמץ נוקשה בטל בששים ורצה לסמוך דמי פירות עם מים לא הוה רק חמץ נוקשה ובששים, ושאלוני חכמי ק"ק הנ"ל אם יש לסמוך על דבריו בזה.

(*Noda B’Yehuda - Orach Chaim Siman* 22)

This question discusses a case involving unleavened bread (*chametz*). Therefore, the “Hannukah” classification was also entirely incorrect.

d. Responsum 7

האמן שאנו עונין אחר ברכת בונה ירושלים אפילו אחר ברכת עצמו אי עונין דווקא בלחישא כדמשמע לישנא דתלמודא או לאו עבדינן אלא כמנהג' דילן דעונין אמן בברכה עצמו כמו שאר אמן:

(*Teshuvot Maharshal Siman* 44)

This question discusses the laws of saying amen after one of the blessings in the grace after meals. Here too, the classification of “Hannukah” is incorrect.

e. Responsum 8

בארצות החום אי שרי לגלגל ביצה על גג רותח כדי שתצלה ונאמ' מאחר שהוא תולדות חמה שרי למיעבד כך בשבת או לא וכן אי אסור להטמינה בחול ובאבק דרכים מבעוד יום או אסור בשבת אפי' לגלגל:

(*Teshuvot Maharshal Siman* 61)

This question relates to rolling an egg on a hot roof on *Shabbat* to roast it (roasting is forbidden on *Shabbat*). The classifier suggested that this question should be classified as “Waking in the Morning” and is incorrect.

f. Responsum 9

אם שכח ע"ה בסעודת פורים אי צריך לחזור או לא:

(*Teshuvot Maharshal Siman* 48)

This question asks whether one who forgot the special addition in Grace after Meals for *Purim* should go back and repeat it. While the question explicitly mentions *Purim*, the addition to Grace after Meals, *al hanissim*, is also said on Hannukah. Therefore, it is reasonable that the classifier assumed that the question was discussing Hannukah although that classification is not the correct one.

Ultimately the SGD classifier remains with an accuracy of 76%. It is interesting to note that of the six completely misclassified responsa, four of them were classified as Hannukah. Further research could be done to try to figure out if this is coincidental or if there is something that can be done to move Hannukah away from being the "universal classification".

II. Logistic Regression

The logistic regression classifier also received a score of 99.36% accuracy on the *Beit Yosef* training data. It received a score of 68.57% on the *Beit Yosef* testing data. When run against the twenty-four responsa, the classifier achieved an accuracy of 58.33%. When running my analysis, I looked at the top three guesses that the classifier made for each of the responsa. Below is a table with the top three guesses for each of the twenty-four responsa as well as the correct answer. If the correct answer was also one of the predictions, the prediction is bolded.

	Prediction #1	***Prediction #2***	***Prediction #3***	***Correct Classification***
1	Shema	**Purim**	Passover	*Purim*
2	**Priestly Blessing**	Shabbat	Prayer	*Priestly Blessing*
3	Passover	**Yom Tov**	Shabbat	*Yom Tov*
4	Shema	Sukkah	**Shabbat**	*Shabbat*

5	Shema	Passover	**Washing the Hands**	*Washing the Hands*
6	Passover	**Chol HaMoed**	Shabbat	*Chol HaMoed*
7	Rosh HaShanah	Yom Tov	**Shabbat**	*Shabbat*
8	Passover	Shabbat	**Yom Tov**	*Yom Tov*
9	The Evening Prayer	**Shema**	Shabbat	*Shema*
10	Shema	Shabbat	**Passover**	*Passover*
11	Passover	Priestly Blessing	**Morning Blessings and Other Blessings**	*Morning Blessings and Other Blessings*
12	Rosh HaShanah	**Sukkah**	The Prayers of Sukkot etc.	*Sukkah*
13	Passover	Shabbat	**Hannukah**	*Hannukah*
14	Shabbat	**Grace After Meals**	Passover	*Grace After Meals*
15	Shabbat	Passover	**Hannukah**	*Hannukah*
16	Passover	**Sukkah**	The Prayers of Sukkot etc.	*Sukkah*
17	Shabbat	Grace After Meals	**Passover**	*Passover*
18	Prayer	Shabbat	**Passover**	*Passover*
19	**Tisha B'Av**	Passover	Shabbat	*Tisha B'Av*
20	Prayer	Passover	**Grace After Meals**	*Grace After Meals*

21	Grace After Meals	Shabbat	**Blessings on Produce**	*Blessings on Produce*
22	Shabbat	**Yom Tov**	Passover	*Yom Tov*
23	Passover	Yom Tov	**Shabbat**	*Shabbat*
24	Passover	Shabbat	**Purim**	*Purim*

Of the twenty-four responsa tested, the prediction with the highest probability was the correct classification for two of the responsa. Eight of the responsa were correctly classified according to the prediction with the second highest probability. The remaining fourteen responsa were correctly classified according to the prediction with the third highest probability. This means that every single one of the responsa was correctly classified within the top three predictions of the classifier. The question then becomes, which classification model is better?

Discussion

I. SGD vs. Logistic Regression

From a pure numbers game, it seems as though the SGD classifier performed better than the logistic regression. At the same time, the logistic regression was able to correctly classify every single one of the test responsa even if it took up to three guesses. However, it is pertinent to analyze the other guesses of the logistic regression classifier and see if the other guesses hold weight or if they are unhelpful. This is especially necessary if the goal is to be able to blindly run the classification and make statements about rabbinic authority as a result.

I looked at the twenty-two responsa that were not classified correctly with the prediction of highest probability to see if the either one or two predictions that had a higher probability were related, loosely related or unrelated to the correct classification. A responsa is considered related

if both (in the case of the correct answer being the third guess) or one (in the case of the correct answer being the second guess) of the guesses could have in theory been used a classification for that specific question or is intrinsically related to the correct topic. For example, one of the questions discussed the *Shema*. The laws of the evening prayer was the earlier guess and much of those laws are intertwined with the laws of *Shema*. Therefore, the guess for that responsa was assumed to be related. A responsa is considered loosely related if the other guesses have some overlap with the topic at hand. For example, the laws of *Shabbat* and the laws of *Yom Tov* have a lot of overlap and it is reasonable that a classifier would incorrectly classify a question about *Shabbat* as *Yom Tov* or vice versa. A responsa is also considered loosely related in a case where the correct answer is the third guess and one of the two earlier guesses is related and the other is unrelated. A responsa is considered unrelated if the other guesses are not connected to the question or the topic in a way that it would make sense for the classifier to have misclassified it in that way.

	Number of Responsa
Related	1
Loosely Related	7
Unrelated	14

It is clear that the majority of the responsa are entirely unrelated to their previous guesses. Thus, although the logistic regression manages to successfully find the correct classification of the forty possible classes within three guesses, at this stage it is not the best model for use in production.

II. Citations vs. Straight Text

When running the tests, I had anticipated that the addition of the citations to the vector would greatly increase the accuracy of my results. Without the citations, the classifiers score as follows: the logistic regression achieves a score of 99.36% on the *Beit Yosef* training data, 72.86% on the *Beit Yosef* test data and 58.33% on the responsa data. As with the citations, the correct answer for each responsa is in the top three guesses. In this instance, four of the responsa were classified correctly with the first guess, six with the second guess and fourteen with the third guess. The SGD classifier achieves a score of 99.04% on the *Beit Yosef* training data, 71.43% on the *Beit Yosef* test data and 58.33% on the responsa data. Fourteen of the responsa were correctly classified. Four of the responsa that were tagged differently from the way they were classified by the classifier, however, are carrying an alternative correct classification. This means that the classifier correctly classified twenty of the twenty-four test responsa. Thus, it seems that if anything, the classifiers perform slightly better without the citations.

This could be for a few reasons.

1. The citations that are being pulled from the text might not be meaningful enough citations. For example, if someone cites the end of a chapter in the Talmud, it might not be an indicator that a certain topic is being discussed as many topics can be covered over the course of a few folios of Talmud.
2. While the CRF model produced great results, the citations being extracted might not be actual citations but instead close enough models that do not serve a purpose in this project.
3. The TF-IDF produces 50,000 features. The citations 10,603. The citations therefore are not numerous enough to outweigh any decisions made on the basis of the TF-IDF features.

III. Comparing Results to ZeroR

While the results of both classifiers seem to be relatively low, it is possible to get a clearer sense of the results by comparing them to a ZeroR baseline.

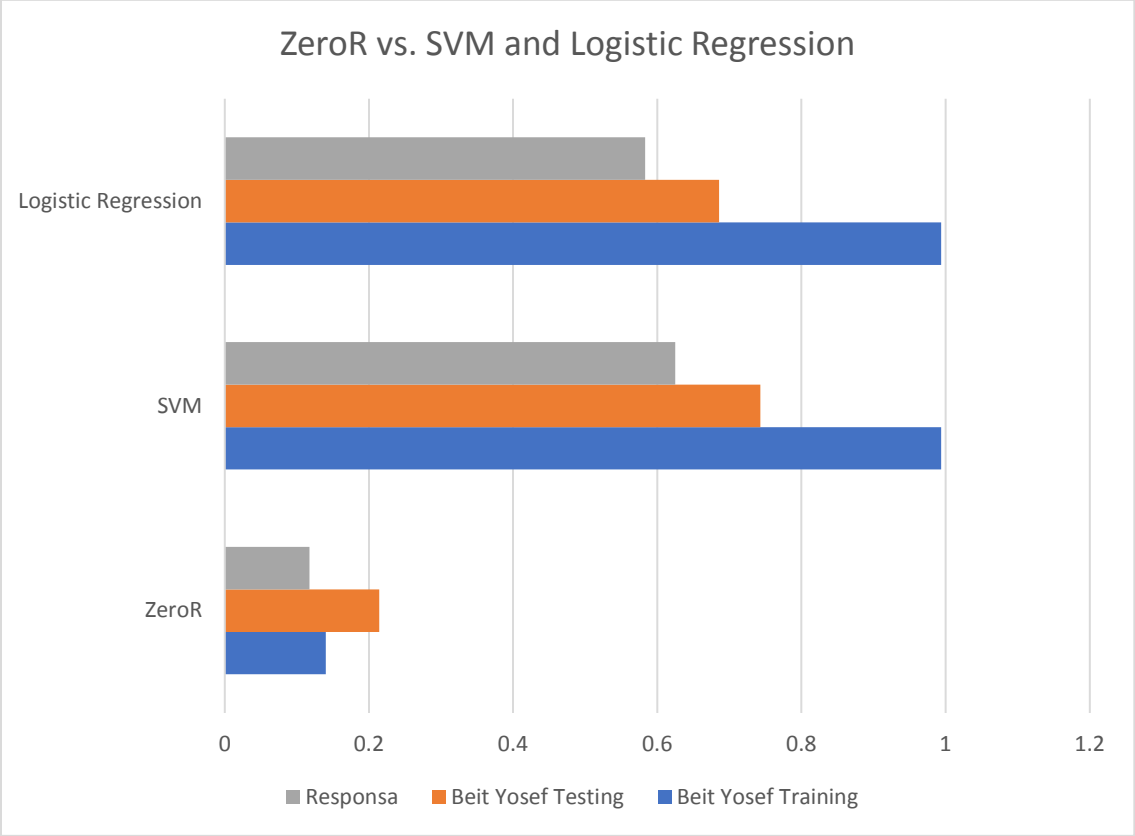

Figure 6: Comparing SVM and Logistic Regression to ZeroR Baseline

Looking at the baseline numbers, it is clear that both the SVM and logistic regression models outperformed the baseline by an extremely large margin. In that sense, although the models need work, they were successful.

Further Research

While the logistic regression model does not produce the best results currently, it provides some advantages for furthering this project. Logistic regression provides the probability of a given answer being correct. A next step, therefore, would be to look at the probabilities of

each of the three guesses and see how close the numbers are to each other. It is possible that for some of the responsa there is a large drop between guess one, two or three and for others, the probability for each of the guesses is relatively similar.

Another area of further research is citations. As stated above, the citations did not produce the anticipated results. There are a few ways in which citations can be improved. The first is to look at the citations alone without any of the TF-IDF features. This would shed light onto whether the citations as they stand are a meaningful way of approaching the problem of categorizing responsa. In addition to this, the citation extractor should be further improved. The list of extracted citations can be analyzed so see what percentage of them are helpful citations and which is any are neutral or potentially harmful. A third way of approaching the citations approach is to expand the citations include actual quotations of text. Perhaps a citation of a folio of Talmud is not enough but if it was accompanied by a key phrase it would better help with classification. Quotations could be utilized in two different ways, the first would be simply to use the quotation as another feature of the responsum. If a rabbi quoted a line from the Rambam, perhaps that line was quoted in the *Beit Yosef* as well. A second way to use quotations is to use the quotation to better refine the nature of a citation. For example, if a rabbi writes, “the Rosh says” and then includes a quotation from the Rosh without a meaningful citation, the quotation could be traced to its source using the Sefaria database.

When analyzing the SGD results, there were a few classifications that were marked as incorrect by the program but were really an alternate correct classification. The program can be adapted to be able to tag text instead of classify it to account for this issue.

Another way in which the project can be taken a step further is by the addition of data. The classification model for citations was based off of tagging only a small number of *simanim*.

If more *simanim* were added to the training set, the citations extracted from the rest of the text would be more meaningful. Additionally, the final model was only tested on twenty-four responsa. Many more responsa could be manually tagged in order to see how the model works on a larger scale.

Once the results of classification of the *Orach Chaim* section of the *Beit Yosef* are high enough, the project to be expanded to include other sections of the *Beit Yosef.*

Conclusion

Classification of responsa is helpful tool to further analyze the rabbinic authority of leading rabbis throughout Jewish history. Although it was initially assumed that a TF-IDF approach would be enhanced through the addition of citation features, the citations do not make much of a difference in the results. Additionally, the SGD classifier performs better than the logistic regression classifier.

Works Cited

Support Vector Machines. Retrieved from https://scikit-learn.org/stable/modules/svm.html#svm

Bird, S., Klein, E., & Loper, E. *Natural language processing with Python*. O' Reily. Retrieved from http://www.nltk.org/book/

Boulton, F. (2018, May 3). Conditional Random Field Tutorial in PyTorch. Retrieved from Boulton, F. (2018, May 3). Conditional Random Field Tutorial in PyTorch. Retrieved from https://towardsdatascience.com/conditional-random-field-tutorial-in-pytorch-ca0d04499463

Brody, S. (2010, July 26). Against the Shulchan Aruch: The Critique of the Maharshal. Retrieved from http://text.rcarabbis.org/against-the-shulchan-aruch-the-critique-of-the-maharshal-by-shlomo-brody/

Cortes, Corinna; Vapnik, Vladimir N. (1995). *"Support-vector networks" (PDF). Machine Learning.* **20** (3): 273–297. CiteSeerX 10.1.1.15.9362. doi:10.1007/BF00994018.

Chen, E. Introduction to Conditional Random Fields. Retrieved from https://blog.echen.me/2012/01/03/introduction-to-conditional-random-fields/

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S.M., Hartmann, S., Gurevych, I., Frank, A. and Biemann, C. (2016): *A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In Proceedings of the LT4DH workshop at COLING 2016, Osaka, Japan*

Karo, Yosef (1923). Beit Yosef – Tur Orach Chaim. Vilna

Lafferty, J., McCallum, A., & Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data (pp. 282–289). Morgan Kaufmann.

Landau, Yehezkel (1880). Noda BeYehuda, Warsaw.

Luria, Shlomo (1574). Teshuvot Maharshal. Lublin.

Mindel, N. (2004, July 1). Rabbi Ezkiel Landau - (5473-5553; 1713-1793). Retrieved from https://www.chabad.org/library/article_cdo/aid/111912/jewish/Rabbi-Ezkiel-Landau.htm

Rajaraman, A.; Ullman, J.D. (2011). "Data Mining" (PDF). *Mining of Massive Datasets*. pp. 1–17. doi:10.1017/CBO9781139058452.002. ISBN 978-1-139-05845-2.

Schorr, M. (2018, May 7). HaMapah. Retrieved from https://blog.hamapah.org/mapping/zoom-in/

Schorr, M. (2018, April 19). Retrieved from https://blog.hamapah.org/mapping/rabbinics-meet-analytics/

Schorr, M. (2019, January 20). Retrieved from https://seforimblog.com/2019/01/who-wrote-the-late-volumes-of-igrot-moshe/

Segal, E. Arba'ah Turim. Retrieved from https://people.ucalgary.ca/~elsegal/TalmudMap/Tur.html