Locust Prediction Analysis
Lily Grier - EPIC Summer Bartlett Fellowship 2020

Background

In this project, I analyzed the monthly country-level locust forecasts produced by the UN Food and Agriculture Organization (FAO). I used natural language processing to extract information from text and find similarities in forecasts and situation reports. I created different criteria for what constituted a correct prediction using varying levels of stringency and changing the number of individual predictions generated from a single forecast report paragraph. I then looked at prediction accuracy trends over time. This project is part of a larger research effort on behalf of the Energy Policy Institute at the University of Chicago that aims to better understand locust activity in hopes of curbing their harmful effects.

Methods

Data Collection and Information Extraction

PDFs of past locust bulletins from 1970 (check year) to May 2020 were scraped from the FAO website using the Python Beautiful Soup package. Each country is listed with a report of the current situation as well as a forecast for the subsequent six weeks. Python package PDFPlumber was used to extract relevant text from the PDF. Due to differences in formatting of the locust bulletins over time, the code only works for reports from July 1997 to present.To incorporate earlier reports into the analysis, code must be written to handle the formatting of those PDFs.

After some cleaning and data wrangling, I created a dataframe where each row represents a single country at one month in time. A row contains a year, a month, the forecast written in that month, as well as that country's corresponding situation reports for the following two months. Two months are included because the forecast extends six weeks into the future. There are about 13,000 rows in the dataframe, which was exported to a CSV file. Due to inconsistent formatting including missing headings and punctuation marks, I was not able to perfectly parse the PDFs.

From this dataframe containing forecast and situation reports, I set out to extract relevant information using spaCy, a free, open-source library for Natural Language Processing in Python. spaCy tagged each word with its part of speech, mapped out grammatical dependencies between words based on sentence structure, and was able to recognize named entities such as dates with reasonable accuracy. Because spaCy does not have a locust-specific module, I used a combination of hard-coding and relying on spaCy's built-in machine learning features to identify the parts of speech and roots of words to identify relevant patterns of text (e.g., one or more adjectives followed by "locusts," "hoppers," "swarms," etc.). Patterns were merged into "entities" and tagged with features such as whether they referred to adults vs. hoppers, solitarious vs. gregarious locusts, and whether the flagged locust group was expected to decline. In addition to locust groups, locust-related actions such as breeding, copulating, hatching, and maturing were flagged as entities. I also flagged locations, both specific (i.e., proper nouns) as well as more general references (e.g., "the southwest coast," "the Algerian border," etc.). While this approach allowed for the speedy extraction of data from over 500 PDFs without needing to read each report and flag pertinent information by hand, it was limited in that spaCy's part-of-speech tagger is not entirely accurate, especially for such a domain-specific project. For this reason, essential locust-related words were hard-coded so as not to miss them through errors in part-of-speech tagging. Errors that occurred several times were corrected through overwriting the model's results, but it is likely some proper nouns, and therefore locations, were likely missed.

Once relevant information was extracted from both forecast and situation reports, the next step was to cross-check that information in order to validate reports. A key challenge in this task was that the report referenced locations and regions by different names. For example, the prediction text might refer to an area using a general term such as "the north," while the corresponding situation might refer to specific cities listed with latitude and longitude. Especially because so many of these general areas are not defined on maps and rely on local conceptions of what the northern and southern parts are that cannot be gleaned by simply bisecting the country along its latitude and longitude. See "Further Directions" for some possible solutions to this issue.

Because predictions for a single country often mentioned several different types of locusts or locust activities occurring in multiple locations, several approaches were taken in evaluating the accuracy of predictions. These approaches varied both at the level of what constituted an individual statement as correct as well as to how overall correctness of predictions for one country in one month was determined. First, pairings of locations and specified locust groups or behaviors were created for each sentence in each prediction. One approach was to group predictions by location, and then if any predicted locust groups or behaviors appeared in the corresponding situation reports for that location, that location's prediction was marked true, and a score was given based on the proportion of locations mentioned in the prediction that had any correct statements. Another approach was to group predictions by sentence, and if any pairing within that sentence was found in the situation report, the sentence was marked as true. Thus, a score was given based on the number of sentences containing true statements. The most stringent approach was to consider each location/locust group pairing individually and create a prediction score based on how many of those pairings were correct without grouping them in any way.

On the individual pairing level, I devised three ways for evaluating correctness. Under the "exact match" method, the words used to describe a locust group at a certain location in the prediction must appear in that location using the same words in the situation report. For example, "mature solitarious adults" will only match to "mature solitarious adults." Under the "general stage" method, locusts are classified by whether or not they are labeled as solitarious (e.g., "isolated," "scattered," "few") or gregarious (e.g., "bands," "swarms," "infestations"). This method also took into account whether locusts were classified as adults or hoppers, but did not differentiate between immature and mature adults. Under these guidelines, "mature swarms" would match with "adult bands," but not with "hopper swarms" or "isolated adults." Finally, under the "any locust" method, any mention of locusts, regardless of life stage or density, matched with any mention of locusts. In all cases, locust behaviors were matched only to words indicating the same action (e.g., "copulating" linked to "breeding" but not to "hatching).

The process of extracting predictions from text and validating those predictions according to the different methodologies is demonstrated in the following tables. This example is from Sudan in July-September 2004.

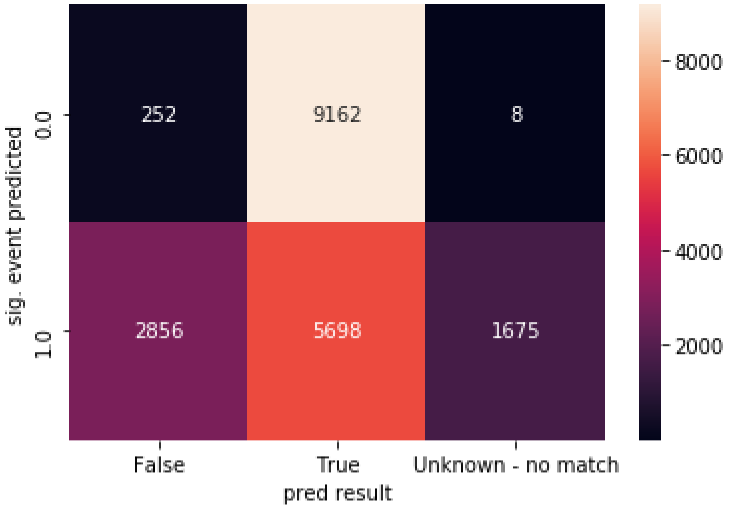
Forecast Text - July	Situation Text 1 - August	Situation Text 2 - Sept
Scattered adults and perhaps a few small groups will appear in the summer breeding areas in Northern Darfur, Northern Kordofan and White Nile States and lay eggs with the onset of the seasonal rains. There is a moderate risk that adult groups and swarms will arrive in Northern Darfur from Northwest Africa. All efforts should be made to monitor the situation in these areas.	**Scattered mature adults** were present at a few places **north of En Nahud (1246N/2828E)** in **Northern Kordofan** on 26-28 August. **No locusts** were seen **north of El Obeid** or in the **White Nile State**. There was an unconfirmed report on 30 August of **swarms** in **Northern Darfur** 48 km from Tine (1501N/2249 E).	During September, **scattered mature adults**, at densities up to 600 adults/ha, were present in **Northern Kordofan**, **north Obeid** (3011N/3010E) and **near Wadi Milk** at 1553N/2808E. **No locusts** were seen in **the north** (**Baiyuda Desert** and **Dongola area**), along the Atbara River and on the **western side of the Red Sea Hills**.
Snippets pulled from text: (scattered adults, Northern Darfur) (scattered adults, Northern Kordofan) (scattered adults, White Nile States) (few small groups, Northern Darfur) (few small groups, Northern Kordofan) (few small groups, White Nile States) (laying, Northern Darfur) (laying, Northern Kordofan) (laying, White Nile States) (adult groups, Northern Darfur) (swarms, Northern Darfur)	Snippets pulled from text: (scattered mature adults, north of En Kahud) (scattered mature adults, Northern Kordofan) (no locusts, north of El Obeid) (no locusts, White Nile State) (swarms, Northern Darfur)	Snippets pulled from text: (scattered mature adults, Northern Kordofan) (scattered mature adults, north Obeid) (scattered mature adults, Wadi Milk) (no locusts, the north) (no locusts, Baiyuda Desert) (no locusts, Dongola area) (no locusts, Atbara River) (no locusts, western side of the Red Sea Hills)

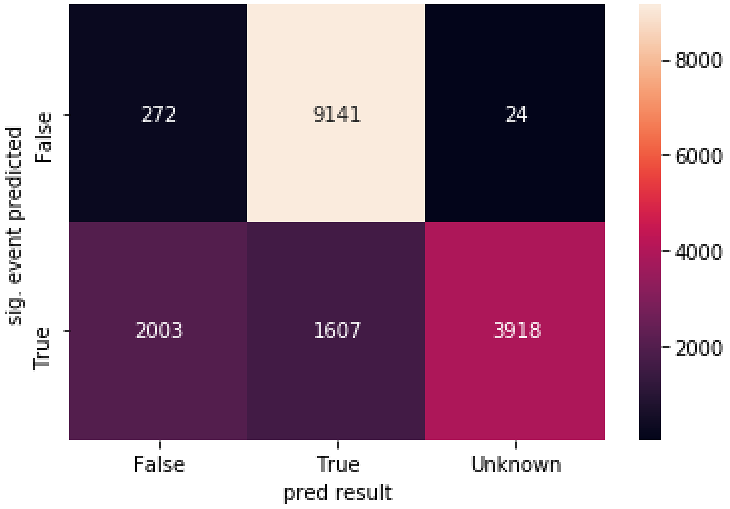
Validation Method	Predictions	Pred Results	Score
Any locusts by place	1. Northern Darfur: locusts 2. Northern Kordofan: locusts 3. White Nile States: locusts	1. True (swarms, Northern Darfur) 2. True (adults, Northern Kordofan) 3. False (no locusts, White Nile States)	**True: 2** **False: 1** **Unknown: 0** **Accuracy: 2/3**
Sentence level by life stage	1. Any of (solitarious adults,	1. True (swarms, Northern Darfur), (adults, Northern	**True: 2** **False: 0** **Unknown: 0**

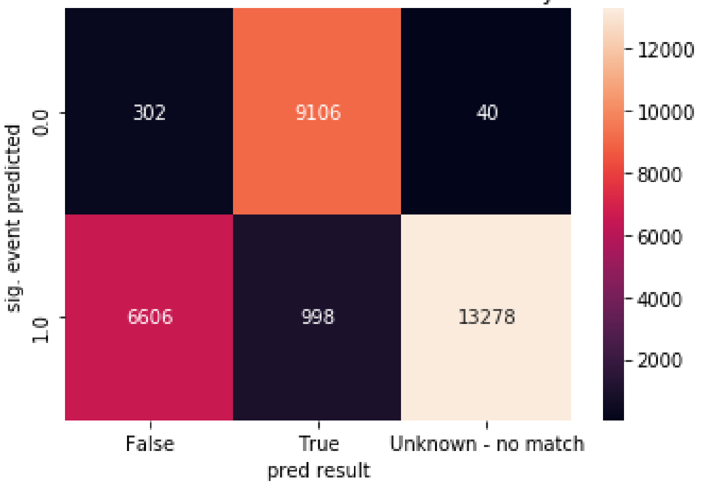
	gregarious adults, laying) mentioned in any of (Northern Darfur, Northern Kordofan, White Nile States) 2. Gregarious adults - Northern Darfur	Kordofan) 2. True (swarms, Northern Darfur)	**Accuracy: 2/2**
Most granular - exact match on locust description only	1. scattered adults - Northern Darfur 2. scattered adults - Northern Kordofan 3. scattered adults - White Nile States 4. few small groups - Northern Darfur 5. few small groups - Northern Kordofan 6. few small groups - White Nile States 7. laying - Northern Darfur 8. laying - Northern Kordofan 9. laying - White Nile States 10. adult groups - Northern Darfur 11. swarms - Northern Darfur	1. False (swarms, Northern Darfur) 2. True (scattered adults, Northern Kordofan) 3. False (no locusts, White Nile States) 4. False (swarms, Northern Darfur) 5. False (scattered adults, Northern Kordofan) 6. False (no locusts, White Nile States) 7. False (laying not mentioned) 8. False (laying not mentioned) 9. False (laying not mentioned) 10. False (swarms, Northern Darfur) 11. True (swarms, Northern Darfur)	**True: 2** **False: 9** **Unknown: 0** **Accuracy: 2/11**

Analysis of Results

The following confusion matrices show the raw counts of the prediction results on the y-axis, and whether or not a prediction was for a significant event (i.e., whether it read anything other than “no significant developments predicted”) on the x-axis. These graphs do not normalize by the number of predictions being higher in the “Most Granular” case, but rather give a general spread of the raw prediction counts.


Any Locusts by Place

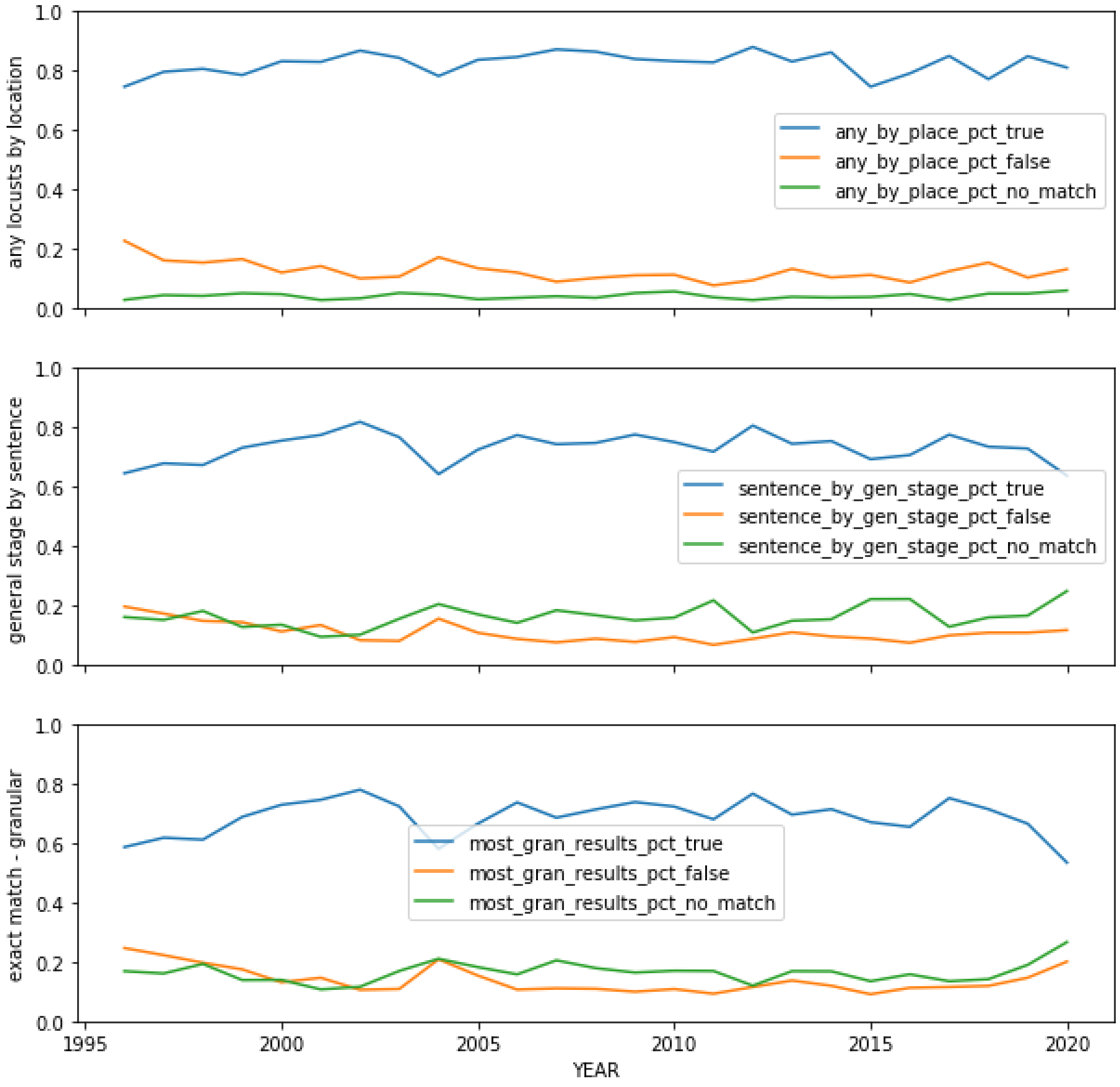

Sentence-Level Accuracy by Life Stage

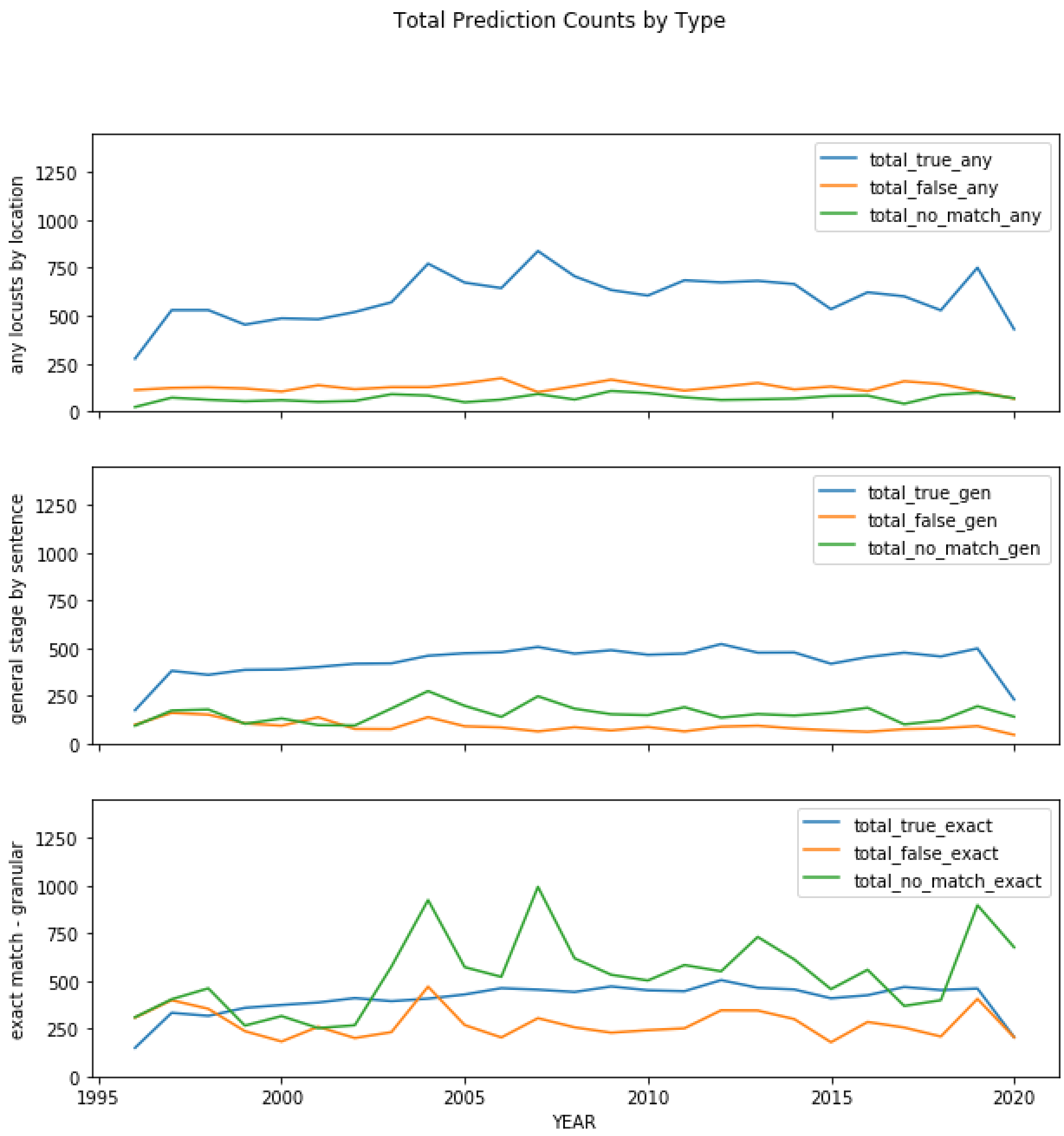

Most Granular Results - Exact Match Only

Consistently across all prediction validation methodologies, we see that the highest number of true predictions occurs when no significant events are predicted and no significant events occur (true negatives). Across all cases, false negatives rarely appear. We see that predictions of significant events are most likely to be correct in the “any locusts by place” scenario than when classifying by sentence/general stage and by exact matches. This suggests the UN has a good sense of where locusts will and will not appear, but is less accurate when it comes to pinpointing specific types of locusts and behaviors (i.e., solitarious vs. gregarious, large vs. small groups, laying vs. hatching).

The following time series show how prediction accuracy changes over time. The first set of subplots depicts raw counts and the second set of subplots depicts the average accuracy score where each score contributing to this average represents one country in one month.

Percent of Total Predictions by Type

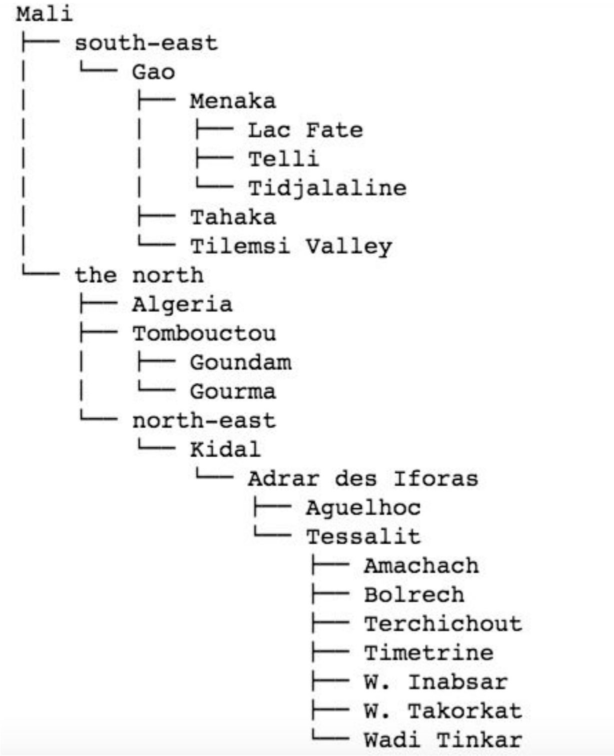




Further Directions

Due to limited time, there are many ways in which I would widen the project's scope, and future work should focus on the following. To enhance the usefulness and accuracy of the prediction validation method I created, location banks should be created for each country to match general terms such as “the north” referenced in the forecasts to specific locales mentioned in the situation reports. To investigate whether this is a worthwhile task, I created a bank of regions for Mali. I was able to match many places that previously had been categorized as unmatched. The process I used to create this example is as follows. Most specific locations are listed in the situation reports with their latitude and longitude coordinates. I used the Google Maps API to search the coordinates and find their administrative regions, then created a tree structure linking labeling the places associated with the coordinates as “children” and the corresponding regions as “parents.” Such a tree structure could also be populated using a database of regions and cities, though the coordinates in the report appear to be more specific. To link vernacular regions such as “the north” and mentions of borders to specific locations I used a combination of searching mentions of regions in encyclopedias and manually looking at maps. I had difficulty finding the full reach of certain areas in Mali such as Tilemsi Valley and Timetrine, which I was able to link to single GPS points in my search but which I suspect extend far beyond those specific points. To do this with complete accuracy, we should consult locals and people with knowledge of informal areas in the country.

Example of Tree Structure for Location Matching



```
Mali
├── south-east
│   └── Gao
│       ├── Menaka
│       │   ├── Lac Fate
│       │   ├── Telli
│       │   └── Tidjalaline
│       ├── Tahaka
│       └── Tilemsi Valley
└── the north
    ├── Algeria
    ├── Tombouctou
    │   ├── Goundam
    │   └── Gourma
    └── north-east
        └── Kidal
            └── Adrar des Iforas
                ├── Aguelhoc
                └── Tessalit
                    ├── Amachach
                    ├── Bolrech
                    ├── Terchichout
                    ├── Timetrine
                    ├── W. Inabsar
                    ├── W. Takorkat
                    └── Wadi Tinkar
```

	False	True	Uncertain
No Match Tree	275	463	86
Match Tree	276	491	57

The table above depicts the “any locusts by place” methodology for Mali with and without using the above tree, which linked cities in two major locust regions, for location verification. Using the tree, we were able to classify an additional 29 predictions, 28 of which turned out to be true and one of which turned out to be false. This suggests that creating a formal way to match locations mentioned by different names could be a worthwhile pursuit in fully understanding the accuracy of these predictions. A csv file of all unmatched locations sorted by country is included in the folder of additional data, should someone wish to take on this task. The frequency with which each unmatched location is mentioned is included in this file such that high-frequency locations can be prioritized in this process.

Furthermore, earlier reports that were scraped from the UN FAO website should be incorporated into the data set. To do this, code would need to be written to handle the formatting of the pre-1997 reports and put them into the csv file, at which point information could be extracted in the same way. This would allow for more meaningful comparisons of prediction accuracy across time. Once these tasks are completed and we have a better understanding of when the predictions are correct, we could use textual analysis tools to try to find patterns in correct predictions and in how predictions and situations change across time. Machine learning techniques could be used to classify the likelihood of locust developments based on features such as past locust activity and whether or not past predictions were correct. The ultimate goal in this would be to create an algorithm that is able to take in a country's historical data and generate accurate forecasts. Such a tool could be invaluable in informing locust control efforts and curbing plagues worldwide.