Analyzing and Classifying United States Agency for International Development Program Evaluations

Ketevan Gallagher

Gabor Computer Systems Research Lab

2023-2024

Background

The United States Agency for International Development (USAID) has long had internal policies that require evaluations of foreign assistance projects. Project evaluations from 1966 to the present are collected on the Development Experience Clearinghouse (DEC), a public digital repository of USAID documents. USAID also maintains the "Evaluations at USAID Dashboard" that contains metadata for evaluations from 2016 onwards.

Evaluation requirements at USAID are determined by both federal law and internal agency policy. In 2011, USAID implemented its biggest change to its evaluation policy in decades. The new evaluation policy dictated that evaluations had to be conducted by a third party, not by USAID staff nor staff of the company that implemented the project on USAID's behalf.

Prior to the policy change in 2011, most evaluations at USAID were commissioned by the organization that implements the program being evaluated. Under the new policy, evaluations are funded directly by USAID, and not the organization that conducts the project. This policy was implemented in an effort to improve the quality and independence of international development evaluations at USAID, and to improve the independence and objectivity of USAID evaluations. It was motivated by a concern that when organizations evaluate themselves, they are more likely to have positive evaluation findings. Additionally, this policy also increased emphasis on impact evaluations. Impact evaluations are those that use rigorous methods to quantitatively estimate the change in an outcome that is due to the program being evaluated. The other category of evaluations is performance evaluations. Performance evaluations are less rigorous, less time intensive, and more often look at whether outcomes occurred but cannot estimate the contribution of the program to the outcome. These evaluations are more qualitative than quantitative and often focus on descriptive qualities such as whether the program was implemented as intended or how the program is viewed by the program participants and beneficiaries.

In this project, my goal was to see if evaluations became more negative over time, as evaluations with less bias are often more negative, which I achieved by assigning evaluations a positive, neutral, and negative score. I also built a machine learning model that can label evaluations as impact or performance evaluations.

Methods

I downloaded all the evaluations from 2000 to the present from the DEC and the Evaluations at USAID dashboard and used the Python package selenium to extract the document ID of each document. To analyze the criticality of each document, I used the Python package VADER to assign each document from the DEC a percent positive, percent negative, and percent neutral score. I then added evaluation specific words to VADER's default lexicon and weighted them heavily to increase emphasis on them and reassigned each document a percent positive, percent negative, and percent neutral score.

To create a model that could determine whether an evaluation is an impact or performance evaluation, I used pyTorch's LSTM layer. I used evaluations from the Evaluations at USAID Dashboard to train my model as they are labeled as impact or performance evaluations. To preprocess these documents, I used the steps outlined in Figure 2. I first tokenized each line so each punctuation mark and word is made into an element in a list. Secondly, I removed all stop words, which are words such as "a" or "the". Thirdly, I removed the words "impact" and "performance". Fourthly, I lemmatized each word and removed any words that contained any symbols outside of the English alphabet. Finally, I used Keras to change each line to a list of numbers. I created my model using pyTorch and used an LSTM layer. By using pyTorch, I could take advantage of TJHSST's GPU servers to increase the speed of my code. I ran my model for 200 epochs.

Figures

Positive Minus Negative Sentiment Scores Created from Lexicon using Added Words for DEC Evaluations Over Time

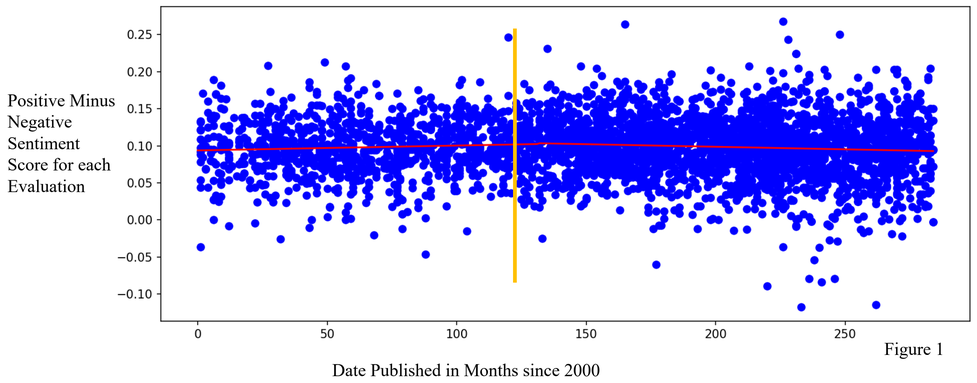

Figure 1

Preprocessing Text Steps

Original Line: "others. while we faced these limitations in evaluating the impact of the farma interventions, we"

1. ['others', '.', 'while', 'we', 'faced', 'these', 'limitations', 'in', 'evaluating', 'the', 'impact', 'of', 'the', 'farma', 'interventions', ',', 'we']
2. ['others', '.', 'faced', 'limitations', 'evaluating', 'impact', 'farma', 'interventions', ',']
3. ['others', '.', 'faced', 'limitations', 'evaluating', 'farma', 'interventions', ',']
4. ['others', '.', 'faced', 'limitation', 'evaluating', 'farma', 'intervention', ',']
5. ['others', 'faced', 'limitation', 'evaluating', 'farma', 'intervention']

Figure 2

Model Results over Time

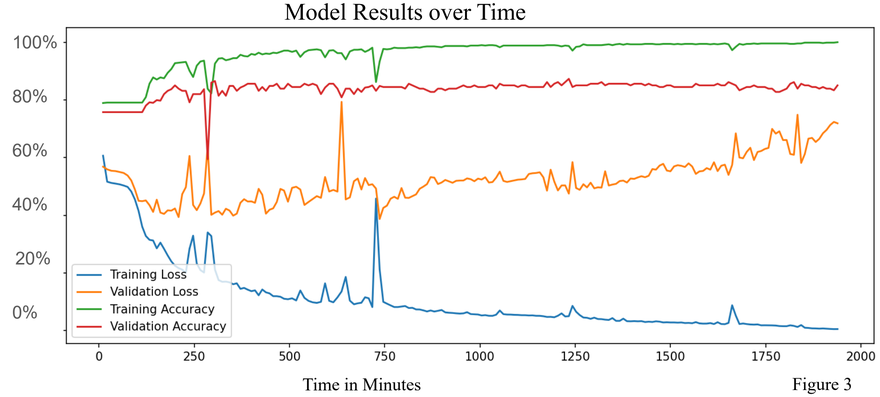

Figure 3

Results

To analyze the results I received when I ran my evaluations through VADER, I conducted a multiple linear regression with an interrupted time series. I used a linear increase over a period of 18 months, starting in January 2011. Although the policy went into effect in January 2011, not all of the evaluations released after this time were in accordance with the new policy. However, after 18 months, most of the evaluations released should have been consistent with the regulations outlined in the policy. In Figure 1, an upward trend in positivity of evaluations can be seen before 2011, but after 2011 there is a downward trend, showing that evaluations became more negative. I used StatsModel to run an Ordinary Least Squares Regression on my results that I received from first using the VADER default lexicon and then the VADER lexicon with evaluation specific words. The p-value for my first regression is 0.001, so when using a standard significance of 0.05, this indicates there was a significant change in sentiment scores after 2011. The p-value for the second regression is 0, which is again significant. Part of the reason the results are significant could be due to the large sample size, so I also calculated an effect size. For both sets of results, my calculated effect size is 1.397, which is considered a large effect size because it is greater than 0.8.

After running my model for 200 epochs, I achieved a training accuracy of 100% and a validation accuracy of 85%. When I used my model to predict the percent of impact evaluations on the DEC, my model predicted that 57.7% of evaluations published before January 2011 are impact evaluations, while 55.9% of evaluations published after July 2012, which is 18 months after January 2011, are impact evaluations.

Conclusion and Future Work

My results show that while the decrease in the positivity of evaluations after 2011 is small because the coefficient for my interactive variable is -0.0001, this decrease is statistically significant.

My model achieved an 85% validation accuracy. I believe this is a successful number considering the size of my training set, which is quite small. Additionally, program evaluations at USAID are not standardized and they are often carried out by different groups within USAID. This leads to a wide variety of formats, which makes it more difficult to differentiate between impact and performance evaluations.

One of the objectives of the policy change was to reduce unwarranted positive findings in evaluations, and a slight but significant decrease in positivity was observed after the change was implemented. This suggests that this part of the policy did have a small effect in the intended direction.

Although my model predicted a decrease in the percentage of impact evaluations after the time the policy went into effect, I believe this is due to an overestimation of impact evaluations. It is expected that less than 25% of evaluations are impact evaluations, which is much lower than the percentage I received. I believe that the reason evaluations after July 2012 have a lower percentage of impact evaluations is because my model predicts modern evaluations more accurately, as I received an 85% validation accuracy for my dataset of evaluations from the Evaluations at USAID Dashboard, which only contains evaluations from 2016 onwards. Overall, due to this high error, I cannot conclude whether the policy had an effect on the number of impact evaluations.

My model seems to have reasonable accuracy for modern evaluations, so a possible extension of this project could include creating a website so USAID employees could upload their evaluations and see if they should be labeled as impact or performance. Evaluations are often mislabeled, so my model could be used to decrease some of the incorrect labeling of evaluations.