Historic Event Discovery with Machine Learning: Final Report

Jason Wei

August 31, 2018

1 Introduction

1.1 Description

This final report for my reading course with Prof. Eugene Santos is a methods report of what I've done over the term in using machine learning to discover historically significant texts. I tried a lot of ideas, and most of them didn't work. My code can be found at https://github.com/jasonwei20/text_discovery_nn.

1.2 Goals

The goal of our study is to use machine learning techniques such as word embeddings, supervised classification, and unsupervised clustering to discover significant (or notable) events in a historical text corpus. The idea is that given a labeled dataset of certain texts, we can train a classifier to rank articles on some spectrum of how much they lean towards their labels, drawing insight from this learned ranking. We hope that this process can be used as empiral evidence to support historical arguments, as well as potentially uncover ideas hidden in the text that are hard to find solely by close readings.

1.3 Dataset

We use the Israeli and Palestinian historical narratives compiled in Side by Side by Sami Adwan as our dataset. The book was purchased in print, and each page was separated (labeled) based on the origin. The book was then sent to a third-party book scanning company for conversion into electronic format. I assign texts from the Israeli narrative the label 0 and texts from the Palestinian narrative 1. Of note, 731 of the 8214 letter sequences (about 9%) in our text corpus were not found in the Common Crawl pre-trained embeddings, indicating a large amount of noise in the conversion of the book from hard-copy to digital.

1.4 Method Overview and Historical Analysis

A text classification model such as a random forest, logistic regression, or feedforward neural net is trained on the labeled dataset described in the previous subsection. After we optimize the model with classifier selection and hyperparameter tuning, we then rerun the model over the entire training set, and get some predicted score for each article. In typical classification tasks, this is done on new data, and an inference is made based on whether the predicted score is greater than or less than 0.5. However, here we use the predicted score to rank each article, and then deem some articles as "notable" based on their predicted confidence and true label. Figure 1 shows a schematic for this training and evaluation process, and Table 1 shows the five notable categories. Not all texts are sorted into one of the five categories.

Predicted Score	True label: Israeli	True label: Palestinian
$(0, \alpha)$	Israeli–biased	Palestinian–sympathetic
$(0.5 - \beta, 0.5 + \beta)$	Neutral	Neutral
$(1 - \alpha, 1)$	Israeli–sympathetic	Palestinian–sympathetic

Table 1: Five categories of notable articles retrieved after re-classifiation of the training set. α is some number less than 0.5 (like 0.2), chosen based on the predicted distribution. β is typically smaller than α (like 0.1), which also depends on the predicted distribution.

Following the sorting of our data into the five categories, it follows that we ask questions regarding what insight these classifications may bring into the controversial history of the Israel-Palestine conflict. A preliminary analysis could involve the following questions:

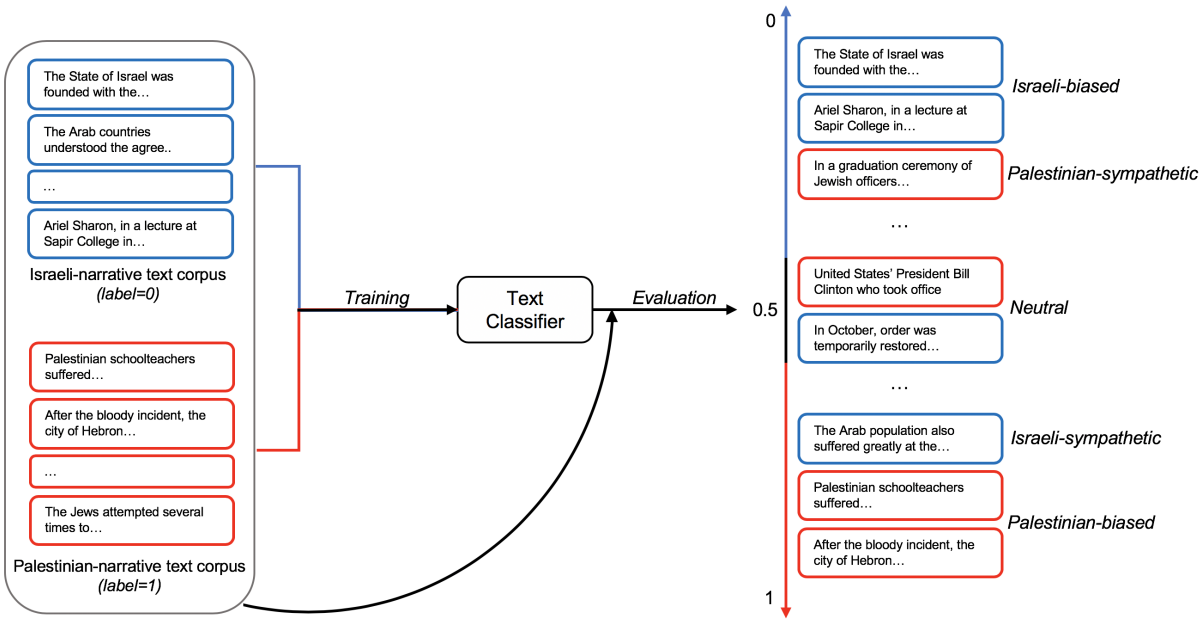


Figure 1: Method overview schematic. The model is trained on a labeled text corpus, and then re-classifies the training set to get some confidence score for each article. Blue borders indicate Israeli sources, while red borders indicate Palestinian sources. Then, misclassifications are labeled into five categories: Israeli-biased, Palestinian-biased, Israeli-sympathetic, Palestinian-sympathetic, and Neutral. Israeli-sympathetic means that the model predicted high Palestinian-bias for an article with a true label of Israeli, indicating that this may have been an event for which the Israeli's were sympathetic to the Palestinian side. Vice versa for Palestinian-sympathetic.

- What events were particularly biased towards either the Israeli or Palestinian narrative? Are these the same events that the Jews and Arabs view so differently?
- Why was the model unable to detect bias for certain events? Was it because the Israelis and Palestinians saw eye-to-eye on some issues? (i.e., did they both hate the British?)
- For what events was one side sympathetic towards the other? Why? Was it because of the undeniable mass suffering experienced during a certain event?

There are several challenges to overcome in this project. On the machine learning side, an accurate classifier must be trained on an extremely small dataset to recognize historical nuance, without directly memorizing the training set. As for historical analysis, the I must either become well-read on the history of the Israeli-Palestinian conflict myself, or I will have to consult with domain experts to make historical inferences from the model's results.

2 Methods

2.1 Wishful Thinking

I started with an idealistic baseline approach that should work well in theory, but in practice ends up with a lot of issues. I split up all the text from each side of the book by paragraph breaks found in the scan, and use each paragraph as a single article, or sample in the data. I end up with something like 1500 data points, and then I train various classifiers on both the bag-of-words and average word embedding representations of the data. I faced several problems:

- For both representations of the data, random forests, logistic regression, and neural nets overfit and memorize the training data with more than 99% accuracy. They end up doing decently on a hold-out development set of 200 samples per class, getting an accuracy of 75%. In typical machine learning problems, where our objective is to classify new texts, having a high training accuracy isn't a problem in itself, as long as the validation accuracy reaches a desirable level. But in our case, since we are rerunning our model over the entire training set, this does become a problem. Imagine a model that perfectly memorizes the training set, and predicts 0.0001 for all Israeli narratives and 0.9999 for all Palestinian narratives. This model tells us nothing new, only information we already know.

- For bag-of-words representations, even weaker classifiers like naive bayes were able to memorize the training set. Their rankings of the articles were heavily skewed to putting longer articles onto the ends of the spectrum, since it was much easier to classify texts that had a lot of words, as there was much more data in them.
- For distributed representations, it was harder to memorize the training set, but still possible for random forests, logistic regression, and a multi-layer neural net over many epochs. The problem here was opposite that of bag-of-words representations: the ranking of articles was skewed towards putting shorter articles towards the ends of the spectrum, since it was more likely that their distributed representations would be more extreme since there were less words. Long articles had a lot of noise, and were more likely to average out over the words, and were almost all into the middle.
- Another thing that I sort of hoped would have happened is some sort of separation in the embeddings based on the categories that I put some of the words into. Before this project I had tried a simple classifier on texts from the bible and quran, and it yielded high accuracy with clear separation of embeddings when visualized with tSNE. However, this was not the case with this problem, which goes to show that detecting bias is a much harder problem than classifying religion. See the first figure in the appendix for how these visualizations turned out. There was almost no separation among categories for the Israeli-Palestinian bias embeddings.

2.2 Troubleshooting

One of the things that I realized is that the validation accuracy is not the only metric that is important for this particular task. In fact, in order to output a spectrum of predictions that can be sorted into useful categories, it was just as important that the classifier was able to predict a range of values, not just values less than 0.01 or greater than 0.99. In fact, maybe I even want to optimize for a distribution of predictions that is close to uniform, not bimodal. I want some classifier that doesn't discriminate based on the length of the article, and also provides some predicted scores with maximum entropy.

- My first approach to solving these issues was to use all the fancy voting, stacking, and hyperparameter tuning methods that I knew. Unfortunately, none of them worked. I first tried to train classifiers on both representations, and then averaging out their predictions by voting and ranking them that way. This brought a marginal increase in the validation set accuracy, but qualitatively, the results didn't seem to be any better. I also tried stacking the representations together. However, even if I scaled the bag-of-words representations to be small, it seemed to still be memorizing them and ignoring the word embeddings. Creating a weaker random forest with less layers and using a smaller neural net simply brought down the validation accuracy to something like 55-60%, which in my opinion was too low to be useful.
- Another thing that I tried, that perhaps was silly on my part but I thought was worth a shot, was using multi-layer neural networks, while at the same time using many dropout layers with high rates. Classic deep learning advice calls for more dropout to combat overfitting, and that is exactly what I tried. I thought that this would also help the fact that my dataset was tiny, so the model wouldn't be able to memorize it; in addition, I had always dreamed of creating my own neural net architecture that was somehow justifiably better than an existing one. So I did a grid search over a wide range of dropout values $(0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8, 0.85, 0.9, 0.93, 0.95, 0.97, 0.99, 0.995, 0.999)$, hidden layers (3–6), and epochs $(10, 20, 30, 50, 100, 200)$, only to find out that none of them really did anything. Small dropout values of less than 0.5 were totally useless, as the model could still easily memorize the training set, for large dropout values, even deeper neural nets would refuse to learn, and everything in between was no better than logistic regression. I ended up spending a lot of time on an idea that was worse than a one-liner in scikit-learn.
- Another one of the things that I tried was removing the most frequent words based on the tf-idf. In typical classification problems, words with high tf-idf are weighted more since they indicate a tendency for certain articles to come from a particular class. However, for this problem I thought it might actually help reduce overfitting, since some words that occur frequently in one corpus may not actually signify any bias. It turns out that this just reduces both the training and validation accuracy by a couple percent.
- I decided to come up with a metric to measure how close a predicted distribution was to the uniform. Basically, chose some n to be the number of buckets, and then calculate the sum of the difference between that bucket and the uniform distribution for that n. The formula ends up looking a lot like mean error:

$$\gamma = \sum_{i=1}^{i=n} \frac{1}{n} \left| x_i - \frac{1}{n} \right|$$

where x_i is the fraction of samples that fall into that particular bucket. Note that γ approaches zero as n approaches infinity, so you can't exactly integrate.

- The thing that ended up solving a lot of these issues, but turns out to be rather unsurprising, is going back and re-cleaning the training data. One of the things that I did is that I actually split up longer paragraphs into smaller sections of two sentences per article. This had a two-fold effect: (1) it increased the number of training samples, and (2) it normalized distribution of article lengths. This both reduced overfitting due to the larger sample size and solved the issue of the model making more confidence predictions based no the length of the article. I guess the saying "garbage in, garbage out" is true. Lesson learned: clean training data as much as possible before even touching any models. After cleaning up the data, I can go ahead and stick to the word embeddings, since they are much faster to train on because of their distributed representations compared to bag of words.
- Another small thing is that in theory, removing stop words should reduce the noise. However, keeping them in improved both the training and validation accuracy by a couple percent. Maybe I should remove them again later, but I will keep them for now.

3 Results

3.1 Classifier Selection

To choose the best classifier, I performed an ablation test with a number of classifiers, and chose the one with the best combination of validation accuracy and maximum entropy. I trained and tested the naive bayes, random forest, logistic regression, and neural network classifiers on the same normalized training set, using both word embeddings and bag of words representations. I use $n = 30$ to get some entropy score γ for each classifier, in addition to measuring its validation accuracy on a hold-out set of 200 samples per class. Figure 2 shows the validation accuracy and entropy of all four classifiers, where I scale entropy to $log(1/\gamma)$, since γ is the area between the predicted distribution and the uniform, so it thus measures the inverse of entropy.

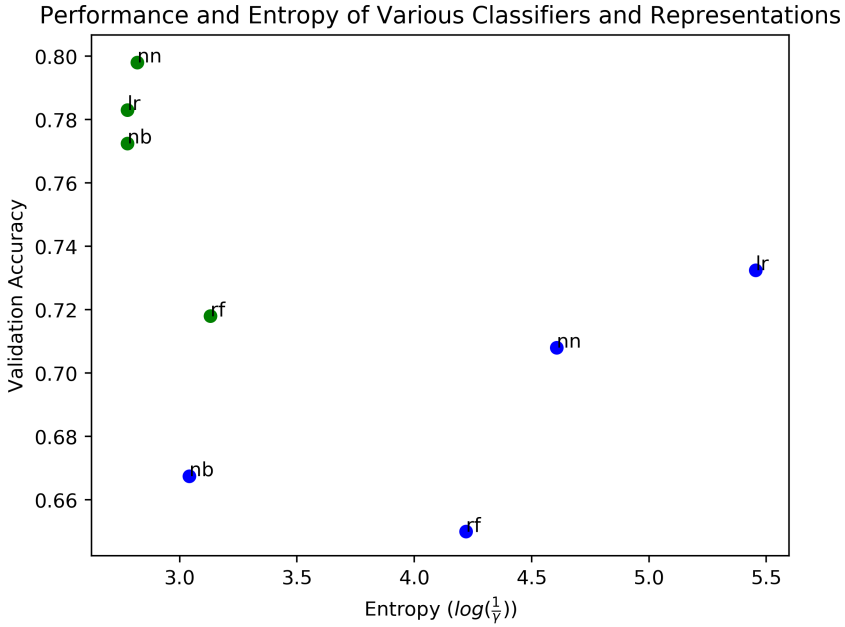


Figure 2: Results of ablation test on four different classifiers and two different numerical representations of articles. Green denotes use of bag of words representations, and blue denotes use of word embeddings. 'nb' = naive bayes, 'rf' = random forest, 'lr' = logistic regression, and 'nn' = neural network.

It becomes apparent that with the bag of words representations, many classifiers can achieve a relatively high accuracy on the validation set. What is not shown on this graph it that this is due to the fact that these classifiers are able to simply memorize the training set, scoring higher than 99% accuracy on the training set. However, this is reflected in the relatively low entropy scores for all of these classifiers; they do not provide us any new information, since all they have done is memorized the training set in a situation where the ground truth labels are known. In selecting our final classifier, we want the one that has the highest performance, but

also provides the most entropy, or the widest spectrum of predictions. This will be the point closest to the upper right corner. We chose the logistic regression model on the word embeddings, since it has the highest validation accuracy with a reasonable entropy. The predicted distribution of this classifier is shown in Figure 3.

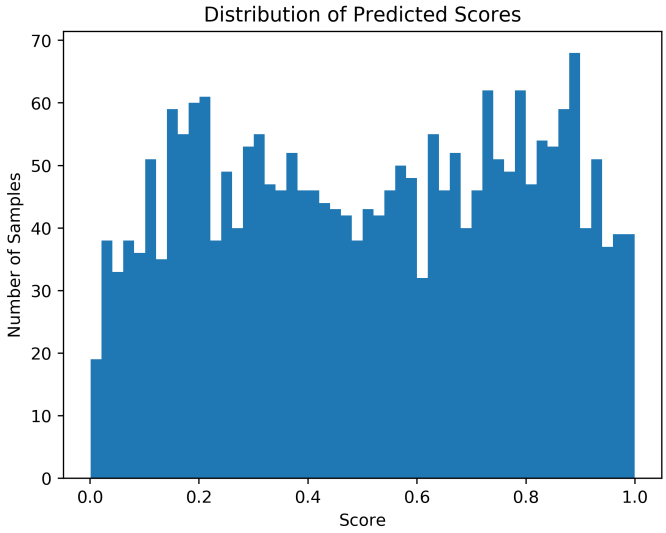


Figure 3: Distribution of predicted scores for the logistic regression classifier on the word embeddings representation. It has the highest entropy of all the classifiers, and looks roughly uniform.

3.2 Validation on External Test Set

To further test my trained classifier to see if it works, I manually collected more data for an external set of different origin to see whether my classifier is able to detect bias. I googled for biased articles of the Israeli-Palestine conflict, and I found one website on algeminer.com [link] titled "New York Times Unleashes Onslaught of Five Op-eds hostile to Israel." This seems like a pretty promising source of Palestinian-biased articles (or at least from the Palestinian narrative), so I went ahead and downloaded them, cleaned them, and put them through the classifier. My classifier detected an average score of 0.534, with 37 articles of Israeli bias and 47 articles of Palestinian bias, which is overall slightly Palestinian-biased, the result we are looking for. However, this is not as strong as I would have liked. I manually examined some of the predicted Israeli-biased articles with high confidence, and it seems that these are often related to President Trump or Iran's nuclear development and foreign policy. The top 5 most confidence results for Palestinian-bias and Israeli-bias, with their confidence scores, are shown in the Appendix.

4 Moving Forward

There is a lot more work to do moving forward. An incomplete list:

1. Working with a historian to find out whether the model actually creates any new insights. I'm still working on getting through the book, but I'm confident that my own study of the text, in collaboration with a Middle East Studies historian, will be able to yield interesting analysis. This is the most immediate work that will allow us to analyze whether our model creates legitimate insight.

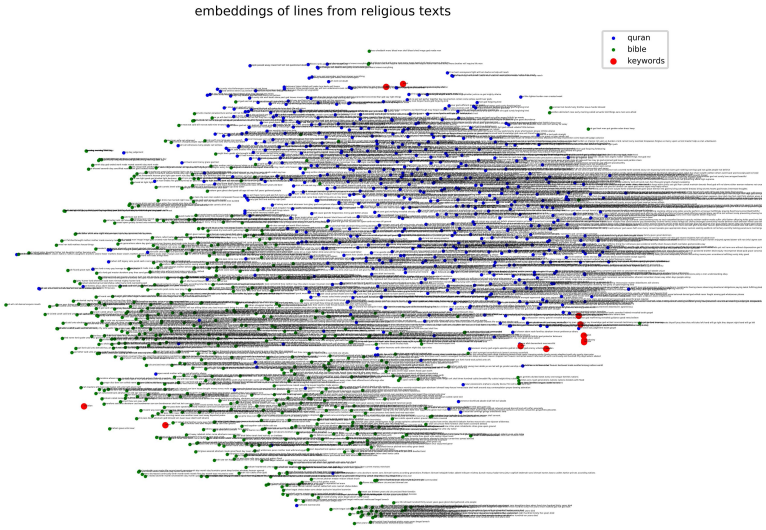
2. It will probably be helpful for me to collect more data for this task. This will involve web-scraping and probably manual collection of data as well.

3. One of the things about the current method of averaging the word embeddings in an article is that it doesn't account for the order of the words. I could use a LSTM RNN to do it, and I plan to do so in the future.

4. I'd like to find a way to use word vector embeddings to represent certain ideas, and then take the dot-product of that vector and another vector to see if certain ideas align. Not sure how to do this with convincing results yet though...

5 Appendix

5.1 Figures



(a)

(b)

Figure 4: tSNE visualizations of embeddings from the bible and quran (a) versus Israeli-Palestinian conflict (b). Obviously, detecting religion is much easier than detecting historical bias.

5.2 Samples of bias classification

The following samples are from a selection of New York Times op-eds that are from Palestinian authors.

5.2.1 Palestinian Bias

1. 0.9883698876510936 raja shehadeh is a lawyer and the author of the forthcoming where the line is drawn a tale of crossings friendships and fifty years of occupation in israelpalestine this essay was adapted from the forthcoming anthology kingdom of olives and ash writers confront the occupation

2. 0.8786440502052348 the raison dtre of the palestinian authority today is not to liberate palestine it is to keep palestinians silent and quash dissent while israel steals land demolishes palestinian homes and builds and expands settlements instead of becoming a sovereign state the palestinian authority has become a protopolice state a virtual dictatorship endorsed and funded by the international community

3. 0.8623526128969461 by dismantling the palestinian authority and reforming the plo the real will of palestinians will be heard whether the endgame is two states or one state it is up to this generation of palestinians to decide

4. 0.8585440334594793 to remove this noose that has been choking palestinians the authority must be replaced with the sort of communitybased decision making that predated the bodys establishment and we must reform our main political body the palestine liberation organization which mr

5. 0.8394813007698414 a security crackdown in saudi arabia before mr trumps visit as well as the bahraini regimes deadly attack on a sitin immediately afterward suggest that the regions despots feel that theyve been given carte blanche to stamp out peaceful dissent

5.2.2 Israeli Bias

1. 0.16165461356740296 we use cookies and similar technologies to recognize your repeat visits and preferences as well as to measure the effectiveness of campaigns and analyze traffic to learn more about cookies including how to disable them view our cookie policy by clicking i accept or x on this banner or using our site you consent to the use of cookies unless you have disabled them

2. 0.15531038105880535 not long ago he drove me to the airport i was going to london for a week and my flight was at in the afternoon twenty years ago the drive took minutes now with so many checkpoints on the way i left the house at noon five hours before the flight

3. 0.14888511483426728 that was precisely the situation in and faced with the realization that the united states sanctions policy was more likely to lead to war than to irans capitulation president barack obama decided to double down on finding a diplomatic solution through secret talks held in oman this time around the american president wont have a diplomatic exit ramp

4. 0.11767123612585151 irans stance on the gaza war is a case in point tehran remained relatively silent and did little to add fuel to the fire compared with what it might have done under other circumstances the iranians understood that they could not secure and sustain a nuclear deal with the united states without shifting their posture on israel

5. 0.10583636679755806 the administration has now said it will conduct a day review of whether lifting sanctions as required by the nuclear deal will be in line with american national security interests but that timeline is not long enough to save the deal and stop the united states and iran from sliding dangerously back to a path toward war