Author Attribution and RNN Text Generation for the Federalist Papers

Elena Ramlow
Analytics, Georgetown University

Abstract

The Federalist Papers are a well-known corpus in the field of computational linguistics, having been the subject of a revolutionary paper on author attribution using Bayesian statistics. This project replicates three previously studied author attribution models—Naïve Bayes, K-Nearest Neighbor, and SVM—on the Federalist Papers and then trains a recurrent neural network using long short-term memory layers to generate text. The RNN models are analyzed using held out validation data to calculate accuracy and perplexity and the generated text is fed back into the author attribution models for author prediction. RNN models are trained on the entire corpus, exclusively on papers written by Hamilton, papers written by Madison, and a combination of the papers attributed to Madison and the disputed papers.

1 Introduction

A wide variety of author attribution models have been trained on the Federalist Papers, after it was a study on the authorship of the twelve disputed papers, that initialized a computational approach to author attribution. However, as technology continues to advance and new revolutions are occurring in computational linguistics, no study to date exists on the ability to train text generating recurrent neural networks using the Federalist Papers.

The purpose of this study is to replicate previously tested author attribution models on the Federalist papers then use recurrent neural networks to generate text, which can then be sent back to the author attribution models for predicting their authorship.

Given the rise in "fake news" and use of author attribution models to detect ghost writers, one can plausibly imagine if given the machine power to do so, an adversarial using recurrent neural networks to generate and falsely attribute text to an author, with malicious intentions. If author attribution models do in fact predict the RNN-generated text to the author the model was trained on, there could be potential negative consequences for society.

The rest of the paper is as follows: first a background on the Federalist Papers and author attribution models used to predict the disputed papers authorship are introduced. Then the methods for producing the aforementioned models are detailed. The results of the models and overall study are discussed. Finally, a general discussion on the overall study and applications to the field are considered.

2 Background

The Federalist Papers are a collection of essays written to the people of the state of New York outlining arguments for ratifying the Constitution. They were written by Alexander Hamilton, James Madison, and John Jay, but signed under the pen name Publius because the authors were also authors of the Constitution, meaning their actual names would bias the reception of the papers. It is historically accepted that Hamilton wrote 51 of the papers, Madison 14, Jay 5, and 3 were a collaboration of Hamilton and Madison. However, the remaining twelve papers have come to be known as the disputed papers. John Jay fell ill after writing his five papers, so the debate over the twelve papers is between Hamilton and Madison.

This author attribution problem was considered "solved" in 1964 by Mosteller and Wallace. Using Bayesian statistics, Mosteller and Wallace (1964) identified 70 function words that could be used to predict the authorship of the disputed Federalist Papers. Their results showed that Madison was the author of all twelve and this has been considered the standard to compare against in subsequent studies of author attribution on the Federalist Papers.

The results of Mosteller and Wallace (1964) revolutionized the field of author attribution, contributing significantly to the field of computational linguistics as a whole, by applying

mathematical properties to the problem. Because of their success with the Federalist Papers, the papers are a well-known corpus to work with when applying various author attribution models. Mosteller and Wallace (1964) are credited with creating the computational field of author attribution and application of statistical models to making attribution predictions.

Further studies into author attribution of the Federalist Papers have made use of a wide variety of supervised and unsupervised learning tasks. Bosch and Smith (1988) used the original function words from Mosteller and Wallace (1964) to create hyperplanes comprised of one to three of the function words. They found that the documents could all be correctly attributed to Madison using only three of the function words. Separating hyperplanes are the underlying mechanism used in Support Vector Machine models.

Both Mosteller and Wallace (1964) and Bosch and Smith (1988) used additional documents written by Hamilton and Madison. Savoy (2013) applied an SVM model to the problem in a similar manner as Bosch and Smith (1998), but unlike the previous two did not use any documents except for the Federalist Papers themselves. Savoy (2013) did, however, utilize additional function words—344 terms defined in previous author attribution research on the Federalist paper. In addition, Savoy (2013) performed TF-IDF vectorization on the words, which ascribes weights to each word. Using these terms and TF-IDF vectorization, the SVM model was able to correctly predict 10 of the twelve disputed papers.

Savoy (2013) also implemented a Naïve Bayes model, again using the 344 terms. This model was able to correctly attribute all twelve of the disputed papers to Madison. Additionally, Savoy (2013) used Delta, Chi-square, KLD, and Z-score methods to predict the authorship. However, only the results of the Naïve Bayes and SVM are relevant to this study.

There does not seem to exist a study of author attribution on the Federalist Papers that implements a K-Nearest Neighbor model. Since KNN models are a commonly used classification method, one will be analyzed in this study.

3 Method

The Federalist Papers are read in and preprocessed then the disputed papers analyzed using three author attribution models, which are a Naïve Bayes model, a K-Nearest Neighbor model, and a Support Vector Machine Model, trained on the known papers. In a similar fashion, recurrent neural networks are trained using the known papers, first using all of the papers, then only those written by Hamilton, those only written by Madison, and finally those written by Madison combined with the disputed papers. The RNN models are used to generate text, which is then fed back into the author attribution models and the authorship of each is predicted.

3.1 Data

The Federalist Papers text is converted into feature vectors for analysis in the author attribution models using the 70 function words outlined by Mosteller and Wallace (1964). While previous studies utilize a wider corpus of training data in addition to the Federalist Papers and function word lists upwards of 300, for simplicity only the papers themselves and the original defined function words are used. Given the relatively small number of papers written by John Jay—only five—and the historical knowledge that the disputed papers were written by either Hamilton or Madison, his papers are excluded from analyses, except when training a recurrent neural network on all of the papers in an attempt to produce the best text generating model possible on the Federalist Papers. Similarly, the papers known to have been collaborated on by Hamilton and Madison are excluded, as they clearly would introduce noise into the binary authorship attribution models.

3.2 Models

For author attribution, three models are trained. The first is a Naïve Bayes model using all 70 of the function words outlined by Mosteller and Wallace (1964). Second is a K-Nearest Neighbor model, created using the brute force algorithm, Euclidian distance, and K = 2. Finally, a Support Vector Machine model is trained using the three function words Bosch and Smith (1998) were able to achieve perfect testing classification with—as, or, and upon. All models are trained using the known papers and tested on the disputed.

A general recurrent neural network model is set up with no pre-trained embeddings and an output embedding size of 100. Three LSTM layers, each with dropout and recurrent dropout equal to 0.1 and 256 memory units, followed by a dense layer with Relu activation, a 0.5 dropout layer, and a dense

layer with softmax activation make up the remaining layers of the model. The papers are converted into sequences of integers with 50 integers/words making up the data and the 51st making up the label. The sequences move along the papers word by word, creating the training data. The model is fit with a batch size of 2048 and 50 epochs. Held out validation data is used to compute accuracy and perplexity.

3.3 Metrics

The author attribution models are all analyzed on their accuracy in predicting the correct author on training and testing data. Measuring both ensures that any signs of overfitting with be realized. Analyzing the quality of a recurrent neural network, however, is more complex. In order to measure the accuracy and quality of the RNN models 20% of the data is held out as validation data in training. This validation data can then be used to calculate an accuracy of the model correctly predicting the next word in the sequence and to calculate perplexity. Perplexity is a commonly used metric for analyzing language models and is equal to two to the power of the log loss of the model, calculated by cross entropy. The perplexity metric can be understood as the number of equally probable events to be chosen from. What this means is that for any given word, the model has the same chance of predicting that word as choosing randomly from the number of items equal to the perplexity. The readability of the text generated by the RNN models is also important, although requires human analysis. If the text follows the rules of English grammar, sentence structure, and could be a plausible (i.e. makes sense) English sentence then it would be considered readable.

4 Results

None of the author attribution models produced results perfectly consistent with the accepted standard, which is attributing all papers to Madison. This is likely due to the small corpus size by not including additional work outside of the papers themselves and utilizing the original function words identified by Mosteller and Rosch (1964). Results could likely be improved by either training the models using additional writing from the authors or utilizing a technique such as TF-IDF to identify function words.

The Naïve Bayes model performed the best, achieving 100% accuracy on the training data and predicting that eleven of the twelve disputed papers were written by Madison. The perfect accuracy on the training data could be indicative overfitting; however, the high accuracy on the testing data suggests that it is a good model. It did not quite achieve the perfect accuracy on the disputed papers as did previous models, which can likely be attributed to the limitation to only the papers themselves as training data and function words used.

The SVM model achieved 98% accuracy on the training data and produced the second highest accuracy on the testing data, at 75%. Though this is lower than previous studies that utilized larger corpora or more function words, it is a decently good model.

The K-Nearest Neighbor model performed poorly on the data. This could be due to the discrepancy in number of papers for each author. With only 14 papers written by Madison and 51 by Hamilton, the nearest neighbor approach is more likely in general to find a Hamilton paper as a neighbor. To produce reliable results K had to be set equal to 2, which is extremely small and likely lead to high variability. Finally, the K-Nearest Neighbor model had 100% accuracy on the training data. This combined with the poor results on the testing data suggests that the model overfit the data. With only $K = 2$, there are no parameters to adjust to account for potential overfitting.

The quality of the recurrent neural networks is assessed using held out validation data. The accuracy of predicted word is calculated. For each of the models the accuracy ranges between 7-8%. Perplexity is also calculated for each and results in a perplexity of between 16,000 and 16,600 per word. Given that the size of the vocabulary of all 85 papers is only 9,911 words, this is a very poor score.

The poor quality of the RNNs can also be seen in the text they generate. The text fails to adhere to English grammar rules, often repeating both words and punctuation. Similarly, it does not follow sentence structure. The text itself is generally unreadable and nonsensical and does not come close to mimicking actual human writing.

5 Discussion

The Federalist Papers is an interesting corpus to work with given its history and role in

computational linguistics. A wealth of data exists on the implementation of various author attribution models on the papers. However, to date, there does not seem to be any published work on training recurrent neural networks on the papers. This is likely due to the small size of the corpus, especially when broken down by author, as is apparent in the resulting generated texts from the RNN models trained in this project.

The authorship attribution models did not perform as well as in previous research, which can be attributed to the limiting of the corpus to the Federalist Papers themselves and using a relatively small list of predefined function words.

What is perhaps the most interesting result of the study is the perfect prediction of the generated texts back to the author they were trained on by the Naïve Bayes model. Whether this truly reflects the strength of the RNNs and Naïve Bayes model is inconclusive, given the poor performance of the RNNs.

6 Conclusion

The combination of small corpus and limited resources for training recurrent neural networks resulted in a model that output nonsensical text riddled with grammatical errors and repetition. Likely increasing both the size of the training text and multiple variables in the model including number of LSTM layers, epochs, and output size of the embedding layer would be necessary to generate text that resembled the Federalist Papers.

Although the RNN models were subpar, the ability of the Naïve Bayes model in particular to correctly attribute each of the generated texts to the author it was trained on. Simplifying both, the underlying mechanisms rely on word frequency; Therefore, the ability of a recurrent neural network to reproduce word frequencies mirroring that of the author it is trained on is a logical conclusion. The implications of this could result in a sort of cyber identity theft if one were to train a model on a writer's work and attempt to pass it off as legitimate.

References

Bosch, R. A. & Smith, J. A. 1988. Separating hyperplanes and the authorship of the disputed Federalist Papers. The American Mathematical Monthly. 105(7): 601-608

Mosteller, F. & Wallace, D. L. 1963. Inference in an authorship problem. Journal of the American Statistical ASSOCIATION. 58(302): 275-309. Doi: 10.2307/2283270.

Savoy, J. 2015. The Federalist Papers revisited: A collaborative attribution scheme. Proceedings of the… ASIS Annual Meeting. 50(1): 1-8. Doi: 10.1002/meet.14505001036.