Explore Machine Learning-Powered Literary Analysis with Text Classification

Qiana Yang
Northwestern University
qianyang2021@u.northwestern.edu

Abstract

The bulk of NLP research focuses on creating and fine tuning language models to perform practical tasks. Common applications include text classification, machine translation, summarization, information retrieval, etc. These tasks rely on models trained on large corpuses of text available through news, reviews, social media posts, etc. Although the performance and integrity of these models depend on the training data, there is less research on the data itself within the NLP community. Traditional literary analysis familiar to linguists and humanities researchers is viewed as a separate, independent domain. I believe a potential new research area that marries literary analysis with a machine learning-powered approach can benefit both disciplines. Traditional literary analysis can deepen our understanding of the training data, thereby furthering our knowledge of the interpretative and predictive potential of language models. Machine learning can in turn help literature researchers discover and verbalize patterns that are not immediately evident, justifiable, or easy to formulate. This work aims to explore the potential of ML-powered literary analysis by using the classification results of a pretrained language model to inform the literary styles in Barack and Michelle Obamas' autobiographies.

1. Introduction

NLP is a burgeoning area of research that aims to build intelligent machines that are able to extract meaning from human text. Literary analysis as a thousand-years-old humanities discipline focuses on qualitative assessment of abstract entities validated by prior disciplinary consensus. One of the challenges when applying literary theories in practical settings is the difficulty of producing quantifiable arguments for decision making. When using text data to solve concrete problems (e.g. to aid our understanding of language bias), however, we hope to have measurable definitions. An example of this challenge is the task of automated bias detection—it is impossible to create a model that detects bias without labeling what constitutes bias.

In cases where quantifiable definitions of social constructs are crucial for NLP tasks, a combination of traditional literary analysis and an ML-powered approach can help. Since current state-of-the-art models are trained on large corpuses of publicly available text, a deeper awareness of the peculiarities of the training data via literary analysis is crucial for interpreting and reasoning with the output of blackbox models. In the meantime, the output of a fairly unbiased blackbox model can accurately reflect the patterns and biases present in the training set. In this work, I hope to provide an experiment that uses a fine-tuned language model to highlight the differences of literary styles in the training data.

The particular texts of my interest are Barack and Michelle Obamas' memoirs. The former President and First Lady have similar morals and values but different experiences, perspectives and aspirations, and these similarities and differences are reflected in *A Promised Land* (Obama, 2020) and *Becoming* (Obama, 2018) respectively. Through text classification, I will use a fine-tuned BERT model to identify and analyze these differences.

2. Related Work

Although there is no direct precedent for this work, I was inspired by scholarship on bias measurement in training data for NLP tasks. A recent paper by Bagga and Piper (2020) outlines an approach to measure and mitigate the effect of bias on literary text classification tasks. Based on their experiments, they conclude that BERT and SVM are robust against mild cases of bias in training data (e.g. gender imbalance).

I also browsed Li et al. 's survey on text classification (2020), from vanilla CNN models to BERT. Since the purpose of this paper is not to explore different text classification techniques but rather to understand how NLP models reflect and act upon stylistic details in training data, I hope not to dwell too much on scholarship on specific modeling techniques.

Finally, I surveyed Hovy and Prabhumoye's paper on sources of bias and counter-measures in NLP applications (2021). Five distinct sources are outlined: bias from data (e.g. selection bias), bias from annotations (e.g. label bias), bias from input representations (e.g. biased word embeddings), bias from models (e.g. underfitting), and bias from research design (e.g. underexposure to certain cultures and languages). Though not directly relevant to the approach in my work, it gives me additional confidence in my belief that understanding training data will be a crucial piece in NLP research.

3. Dataset

The data used in this work consists of Barack Obama's memoir *A Promised Land* and Michelle Obama's memoir *Becoming*. The data is preprocessed in a way that only preserves paragraphs greater than 100 characters (including spaces). This ensures that shorter sentences, usually comprised of standalone dialogues, brief remarks and exclamations, are omitted from the training process. The preprocessed data is saved as text files in the data folder in the project repository.

After preprocessing, there is a total of 2831 paragraphs in *A Promised Land* and 1503 paragraphs in *Becoming*. *A Promised Land* has approximately 290K words, 152K of which are content words. *Becoming* has approximately 162K words, 80K of which are content words. Among *A Promised Land*'s most popular words are "like" (596 occurrences), "people" (570 occurrences), time (555 occurrences), "us"/"US" (452 occurrences), "made" (426 occurrences), "house" (426 occurrences), and "president" (411 occurrences). Among *Beloved*'s most popular words are "Barack" (816 occurrences), "us"/"US" (494 occurrences), "like" (381 occurrences), "people" (367 occurrences), "time" (365 occurrences), and "school" (306 occurrences).

4. Method

The goal of this work is to explore how to use NLP models to further our understanding of the training data. My hypothesis is, in order to evaluate and quantify the presence of bias in a piece of text, we may use a fairly unbiased language model capable of capturing semantic and syntactic patterns in unstructured text to perform a simple task, and use the model output to further our understanding of bias in the original text. The specific task in this project is to feed Barack's and Michelle's text into a state-of-the-art classification model, analyze model predictions, and identify literary elements that belong distinctly to Barack and Michelle respectively. I broke the modeling and evaluation process down into three tasks:

1. Exploratory analysis—counting the most frequent words in the source text.

2. Text classification—classifying authorship with a fine-tuned BERT model and analyze where false positives and false negatives occur.

3. Text prediction—predicting authorship of keywords with the model and interpret results.

The first task is completed via tokenization during the preprocessing step. I used NLTK's RegexpTokenizer method to compare Barack's and Michelle's lexica at a high level.

The second and third tasks are completed with the aid of tensorflow and HuggingFace's transformer package. I deliberately used BERT to perform this step because it is robust against many instances of bias according Bagga and Piper's work. The model is capable of capturing the intricate details of unstructured text. It can allow me to confidently conclude that any wrong predictions in the output is probably due to the noise in the training set rather than the model itself. In other words, if I were to use a less complicated model, then it would be unclear whether the wrong predictions is due to bias in the model or bias in the training set.

The final model is initialized with the pretrained "bert-base-uncased" parameters and fine-tuned for an additional 100 epochs on 80% of stratified and randomly sampled paragraphs from *A Promised Land* and *Becoming*. The model is then used to predict on the remaining 20% paragraphs. After evaluation of test set results, it is used to predict on paragraphs, sentences and keywords outside of the memoir corpus.

5. Result

When I first read *Becoming*, I found it a refreshing experience. Contrary to my typical impression of political biographies, Michelle's memoir was candid and unpretentious. Barack's, too, was revealing, but I always felt that there was a detached quality to it, that he was writing for an audience larger than himself. The result of my exploratory analysis and text classification confirmed my hypothesis: while Michelle's stories are more down-to-earth, Barack's prose is more idealistic and ponderous (which exactly matches Michelle's assessment of him in *Becoming*). *A Promised Land* is twice as long as *Becoming*, with nearly twice as many content words. Out of the top 100 content words in *A Promised Land*, most are impersonal entities related to his work in the government (e.g. "president", "campaign", "country", "senate", "political", etc.), few of which end up in Michelle's lexicon. The word "Michelle" ranked as the 31st most popular content word with only 289 occurrences, compared to "Barack" appearing a total of 816 times and ranking as the most popular content word in *Becoming*. The top 100 content words in Michelle's memoir include "home", "life", "kids", "mother", "Malia", "girls", "friends", "father", "parents", and "Sasha", all of which are missing from Barack's top vocabulary. "Family" ranks as the 27th most popular word in *Becoming* and 94th in *A Promised Land*. This preliminary analysis made me to hypothesize that while Barack was more absorbed with his identity as a public figure, Michelle was more concerned with her role in the community and private home—at least on paper.

The result of text classification aligns with this analysis. Even though my goal is not to focus too much on evaluation metrics, the BERT model does perform the classification task admirably, achieving an overall test accuracy score of 85% with the following precision, recall and f1 scores:

	Precision	Recall	F1 Score
Barack	0.87	0.90	0.89
Michelle	0.80	0.75	0.78

Table 1: classification result

On a micro scale, the model is better at identifying Barack's passages than Michelle's. For the purpose of this work, the model is robust enough for us to probe into the intricacies of the training data. Out of the 567 paragraphs by Barack in the test set, 56 are wrongly predicted as Michelle's. Out of these, 49 are non-dialogue passages, 14 of which are about Michelle, kids, parenthood, personal family, and families in general (e.g. military families), 6 of which are about female politicians, in particular Hillary Clinton, and another 6 on community building. Here are some sample passages with their corresponding tags:

Family: *Every parent savors such moments, I suppose, when the world slows down, your strivings get pushed to the back of your mind, and all that matters is that you are present, fully, to witness the miracle of your child growing up. Given all the time I'd missed with the girls over years of campaigning and legislative sessions, I cherished the normal "dad stuff" that much more.*

Female politician: *On the flight to Des Moines, I tried to appreciate the frustrations Hillary must have been feeling. A woman of enormous intelligence, she had toiled, sacrificed, endured public attacks and humiliations, all in service of her husband's career— while also raising a wonderful daughter.*

Community: *And yet they did connect. Arriving in town with a duffel bag or a small suitcase, living in the spare bedroom or basement of some early local supporter, they would spend months getting to know a place—visiting the local barbershop, setting up card tables in front of the grocery store, speaking at the Rotary Club…They worked each day to exhaustion and fought off bouts of loneliness and fear. Month by month, they won people's trust. They were no longer strangers.*

The content of passages in *Becoming* that are wrongly predicted as Barack's is more diverse and wide-ranging. The narrative style is largely matter-of-fact. Even when family and community are mentioned, they are spoken of in a neutral, detached tone. For instance, note the narrative style in the following passage:

Motherhood became my motivator. It dictated my movements, my decisions, the rhythm of every day…Barack and I studied little Malia, taking in the mystery of her rosebud lips, her dark fuzzy head and unfocused gaze, the herky-jerky way she moved her tiny limbs. We bathed and swaddled her and kept her pressed to our chests. We tracked her eating, her hours of sleep, her every gurgle. We analyzed the contents of each soiled diaper as if it might tell us all her secrets.

The classification result shows that differences between Barack's and Michelle's styles and perspectives do exist, and the BERT model is capable of detecting the underlying text patterns. Wrong predictions can and do take place when the content or narrative style of a particular passage becomes more ambiguous.

Finally, I fed custom keywords into the model and observed its predictions. Keywords classified as Barack's include legislation and issue-related terms with a neutral connotation (e.g. "immigrant", "wealth", "law", "healthcare", "income", "inequality", "diplomacy", "war", and "enforcement"), idea-driven actions and entities (e.g. "idea", "idealistic", "dream", "analyze", "think", and "insightful"), abstract concepts (e.g. "value", "morals", "Kant", and "fair"). Keywords classified as Michelle's include all personal pronouns except "they", family-related entities (e.g. "family", "marriage", "Malia", and "Sasha"), positive emotions (e.g. "love", "feel", "passion", "faith", "beautiful", "good", "smart", "cool", "hardworking", and "accomplished"), community-related entities (e.g. "people", "public", "school", and "ghetto"), and practical day-to-day considerations (e.g. "money", "job", "life", "responsibility", and "practical"). The result here reinforces the argument that Michelle's narrative is more rooted in reality whereas Barack's narrative is split between details of his political work and grand philosophizing.

6. Discussion

In this paper, I proposed an approach to marry ML-driven techniques with traditional literary analysis. Overall, I have found that the pretrained BERT model is capable of capturing literary patterns in the training text and highlighting these patterns through the model output. The result of the classification task agrees with and provides quantifiable proof for my empirical assessment of Barack's and Michelle's writing styles. I hope that this experiment can be a useful framework for future interdisciplinary research in the NLP community.

References

Sumyam Bagga and Andrew Piper. 2020. Measuring the Effects of Bias in Training Data for Literary Classification. In *Proceedings of LaTeCH-CLfL.* Barcelona, Spain.

Dink Hovy and Shrimp Prabhumoye. 2021. Five Sources of Bias in Natural Language Processing. In *Language and Linguistics Compass. Vol 15, Issue 8.* https://doi.org/10.1111/lnc3.12432.

Qian Li, et al. 2020. A Survey on Text Classification: From Shallow to Deep Learning. *IEEE Translation on Neural Networks and Learning Systems. Vol 31, No. 11.*

A. Supplementary Material

The link to this paper's GitHub repository is https://github.com/MSIA/qyt9304_msia_text_analytics_2021/tree/project. The repo consists of Barack and Michelle's memoirs in text format, the fine-tuned BERT model, source code for text preprocessing, model training, and a REST API for model deployment.