Deep Causal Inference for Average Treatment Effect Estimation of Poems Popularity

Derek Farren*
Department of Computer Science
Stanford University
dfarren@stanford.edu

Abstract

Poetry is a complex art form. This research is focused on exploring and understanding poetry better by using deep learning. On one hand, there are many different styles of writing a poem, and on the other only a small fraction of poems end up passing the test of time and becoming popular. The objective of this work is to find the causal relationship (if any) between the poem style and it's popularity. In particular, this project answers the following question: Does a specific rhythm in a poem makes it more popular?

1 Introduction

What makes a poem "good"?

The answer ultimately lies with the reader of the poem, but there is a certain consensus as to what makes a poem "good" or "bad."

According to the critic Coleridge, prose is "words in their best order," while poetry is "the best words in their best order.

Poetry demands precision. The novelist can get away with less than precise expression from time to time because the story will pull the reader along. The job of the poet is to create a picture in the mind and an emotion in the heart. Every single word counts. The wrong choice–a word with the wrong connotation or the wrong number of syllables or an unlovely combination of consonant sounds–spoils all.

The underlying thought of the poem is also important. Some poems are written to create a picture only, but the most memorable poems also convey a universal truth about the human condition. For me, a "good" poem leaves me with goosebumps along my arms. I think a poem is "bad" when it lacks a discernible point and sounds like prose.

People are led to write a poem because they have been strongly moved by some event. They've experienced a strong emotion, received an insight, and wish to capture the experience in words. Only a few, however, succeed in turning the experience into a poem that will be meaningful to another person.

On his site dedicated to examples of bad poetry, Prof. Seamus Cooney observes that most bad poetry is "simply weak and ineffectual and lacking in interest."

He says that memorably bad poetry is created by "a poet unaware of his or her defects." He says that a really dreadful poem is the product of "the right combination of lofty ambition, humorless self-confidence, and crass incompetence. . . ." He collects examples of bad poems as a teaching device. Here's an excerpt from one of Prof. Cooney's bad poems:

*Thanks to Thai Pham [3] for his generous help sharing his causal inference expertise.

> Twas the year of 1869, and on the 19th of November, Which the people in Southern Germany will long remember, The great rain-storm which for twenty hours did pour down, That the rivers were overflowed and petty streams all around. –from "Saving a Train" by William McGonagall (1825-1902)

A successful poem does not have to rhyme or scan or have a certain pattern of lines. It does need to paint a picture with carefully chosen words. It should have a point that a reader unknown to the poet can respond to.

Poets can study a wide variety of poetry–good and bad–in order to learn what works and what does not... or they can use deep learning.

2 Related work

This is a poem written by a deep learning model developed at Google Brain:

I want to talk to you.
I want to be with you.
I don't want to be with you.
I don't want to be with you.
she didn't want to be with him.

There has not been much work done in the intersection between poetry an computer science, especially when it comes to deep learning. Google published research showing their ability to generate poems using a RNN [8]. Then [9] followed by doing the same in the chinese languaje.

3 Rhythm and Meter in English Poetry

English poetry employs five basic rhythms of varying stressed (s) and unstressed (w) syllables. The meters are iambs, trochees, spondees, anapests and dactyls. In this document the stressed syllables are marked in boldface type. Each unit of rhythm is called a "foot" of poetry.

The meters with two-syllable feet are:

IAMBIC (ws) : That **time** of **year** thou **mayst** in **me** be**hold**

TROCHAIC (sw): **Tell** me **not** in **mourn**ful **num**bers

SPONDAIC (ss): **Break, break, break**. On thy **cold gray stones,** O **Sea!**

Meters with three-syllable feet are

ANAPESTIC (wws): And the **sound** of a **voice** that is **still**

DACTYLIC (sww): **This** is the **for**est pri**me**val, the **mur**muring **pines** and the **hem**lock (a trochee replaces the final dactyl)

The meter is implicit within the text. For each poem line the percentage of ww – that is, how often two weak syllables appear together – divides a binary from a ternary meter. And the percentage of fourth syllables that are strong divides front from backheadedness. The percentage of ww and of strong fourth syllables are used to guess the poem's meter. A poem that has a large number of ww and a high percentage of fourth syllables that are strong is typically dactylic. A poem that has a low number of ww and a high percentage of syllables that are strong is typically iambic (the most common English foot).

In order to find the meter for each of my poems in the corpus I used prosodic [1].

4 Popular Poems

What defines a Popular poem? That is a hard question to answer. Based on the feedback I got from Mark Andrew Algee-Hewitt [2], a reasonable way to label a poem as popular is if it was published.

In order to find our set of popular poems I matched all poems in out corpus with a published poem by searching for it on the project Gutenberg [5]. If I find the poem, it was published and hence it is popular. If I don't find it, it is not popular.

5 Dataset

My dataset consisted on approximately 600,000 poems generously given to me by the Stanford's Department of English. It's worth mentioning that parsing the data took around half the time taken to finish this project.

6 Causal Inference

This model is based on the research of Thai Pham [3].

$X_i \in \mathbb{R}^p$ is the feature vector of the poem, $W_i \in \{0, 1\}$ is the treatment/intervention variable, in this case the whether the poem is Iambic or not. I choose Iambic because it is the most popular meter in the English language. $Y_i \in \mathbb{R}$ is the outcome of interest, in this case the poem popularity label. We are interested in estimating the average treatment effect (ATE) of W on Y. In particular, we want to estimate τ, where

$$\tau = \mathbb{E}(Y|X, W = 1) - \mathbb{E}(Y|X, W = 0)$$

To this end, we use the Doubly Robust Estimator as a special case of Targeted Maximum Likelihood Estimator [7],[6]. The steps are as follows.

1. Using treated data $\{i : W_i = 1\}$ to estimate $\mu_1(x) = \mathbb{E}[Y|X, W = 1]$ with estimator $\widehat{\mu_1}(x)$
2. Using control data $\{i : W_i = 0\}$ to estimate $\mu_0(x) = \mathbb{E}[Y|X, W = 0]$ with estimator $\widehat{\mu_0}(x)$
3. Using all data to estimate $e(x) = \mathbb{P}(W = 1|X = x)$ with estimator $\widehat{e}(x)$
4. The (Doubly Robust) ATE estimator $\widehat{\tau}$ for τ is given by
$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left[W_i \times \frac{Y_i - \widehat{\mu}(1, X_i)}{\widehat{e}(X_i)} - (1 - W_i) \times \frac{Y_i - \widehat{\mu}(0, X_i)}{1 - \widehat{e}(X_i)} + \widehat{\mu}(1, X_i) - \widehat{\mu}(0, X_i)\right]$$
5. The standard error is estimated by first defining
$$IC_i = W_i \times \frac{Y_i - \widehat{\mu}(1, X_i)}{\widehat{e}(X_i)} - (1 - W_i) \times \frac{Y_i - \widehat{\mu}(0, X_i)}{1 - \widehat{e}(X_i)} + \widehat{\mu}(1, X_i) - \widehat{\mu}(0, X_i) - \widehat{\tau} \quad (1)$$
$$\text{and } \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} IC_i^2. \quad (2)$$
Then the standard error is estimated by $\dfrac{\sigma}{\sqrt{n}}$.
6. The 95% Wald-type Confidence Interval is: $\widehat{\tau} \pm z_{0.975}\frac{\sigma}{\sqrt{n}}$.

In the state of the art research, these estimators are found using linear regression. We will use deep learning.

7 Base Line Model

The base line model uses a multilayer feedforward networks with the parameters shown in Table 1. The feature vector here is the 50 dimensional average word vector using all words in the poem and GloVe word vectors.

The F-1 scores for each of the models are shown in figure 1. Using these models we now look for a causal relationship between meter and poem popularity using the ATE estimator defined in 6.

The ATE estimator had a 95% Wald-type confidence interval of [-5.783, 7.564]. This model does not show a causal relationship between Iambic poems and its popularity.

Table 1: Sample table title

Prameters	
Name	Value
initial learning rate	0.001
number of epochs to run trainer	100
size of the minibatch	500
probability of keeping data after dropout	0.8
n_input	50
n_hidden_1	128
n_hidden_2	85

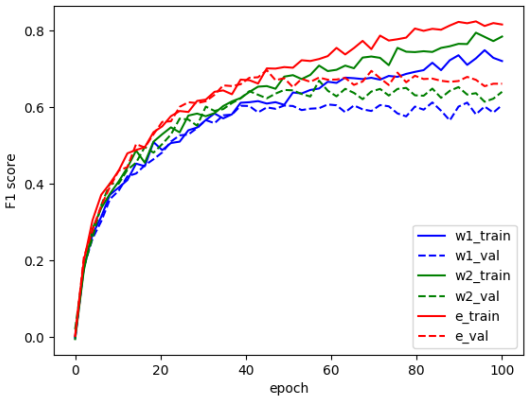

Figure 1: The three models that generate our Base Line ATE estimator have F1 scores in the range [0.5, 0.7]

Table 2: Sample table title

Prameters	
Name	Value
initial learning rate	0.003
number of epochs to run trainer	100
size of the minibatch	500
probability of keeping data after dropout	0.8
n_input	500
n_hidden	256

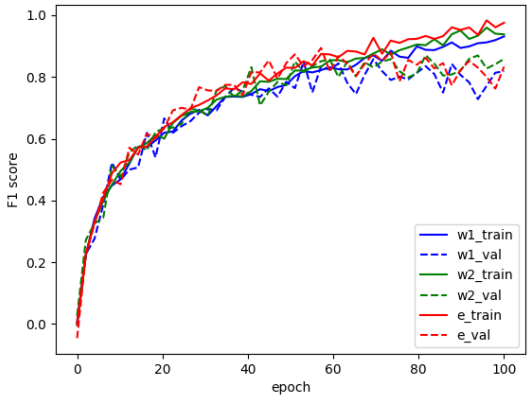

Figure 2: The three models that generate our Base Line ATE estimator have F1 scores in the range [0.7, 0.8]

8 LSTM Model

I compared the Base Model results with the ones obtained by a LSTM model using the parameters shown in Table 2.

This model uses as input the first 10 words of a poem. I tried to use the whole poem, but the running time was too long on Azure's GPUs.

This model performed better. The F1 scores are shown in figure 2, and most importantly the 95% Wald-type confidence interval is [-0.014, 4.395], which shows that a Iambic meter cause a poem to be popular with very high probability.

9 Conclusions

There is strong statistical significance that support the thesis that the Iambic meter causes a poem to be popular. The selection of model used to estimate the parameters of ATE is very important in order to be able to make valid causal inferences. This research used an LSTM but probably there are other deep learning models that can fin causal relationships much more accurately than the commonly used linear regression.As far as I know nobody has tested ATE using deep learning.

Acknowledgments

Thanks to Mark Andrew Algee-Hewitt [2], to Ignacio Cases [4], and especially to Thai Pham [3].

References

[1] Ryan Heuser, Josh Falk, and Arto Anttila. https://github.com/quadrismegistus/prosodic

[2] https://english.stanford.edu/people/mark-algee-hewitt

[3] https://www.gsb.stanford.edu/programs/phd/academic-experience/students/phillip-thai-pham

[4] http://stanford.edu/~cases/

[5] www.gutenberg.org

[6] Susan Athey, Guido Imbens, Thai Pham, and Stefan Wager. Estimating Average Treatment Effects: Supplementary Analyses and Remaining Challenges. *The American Economic Review: Papers and Proceedings*, forthcoming, 2017.

[7] Mark J. Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

[8] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz Samy Bengio Generating Sentences from a Continuous Space

[9] Xiaoyuan Yi, Ruoyu Li, Maosong Sun Generating Chinese Classical Poems with RNN Encoder-Decoder