Statistical Significance Testing in Machine Learning

Spencer Slaton, Leon Weingartner

***Abstract*—We recreate the study of statistical significance testing within the context of neural networks. This research was produced by Dennis Ulmer, Christian Hardmeier, and Jes Frellsen with a paper titled "Deep Significance – Easy and Meaningful Statistical Significance Testing in the Age of Neural Networks" ICLR (2022) *[1]*. The authors comment on the lack of statistically rigorous testing within the rapidly expanding field of deep learning. To rectify this, they provide a Python package for implementing a novel and powerful significance test called *Almost Stochastic Order* (ASO). Here, we will revisit the ASO statistic and compare it to a variety of well-established significance tests across numerous scenarios to illustrate the usefulness of ASO and statistical testing in general under machine learning architectures including a simulation study using convolutional neural networks (CNNs) and multi-layer perceptrons (MLPs).**

I. Introduction

When choosing between two different machine learning architectures, the industry standard uses a single comparison of error scores on a single validation/testing set to determine the superior model. This methodology can prove to be somewhat unreliable as you can produce a lower error score from an objectively worse model. This widely used model selection technique lacks statistical robustness and weakens sequential ML research. There seems to be an industry wide shyness to statistical testing which is an incredibly underused tool that we hope to shed some light on.

The general need for statistical testing stems from the stochastic properties of neural networks that can produce random outlier results. Reimers & Gurevych (2018) *[2]* suggests the single score comparison is insufficient and requires more evidence. This has been a known problem in recent studies involving machine learning. The use of new statistical testing methods to evaluate model performance was proposed in Dror et al. (2018) *[3]*. Our focus in this paper will revolve around the powerful Almost Stochastic Order (ASO) test outlined in greater detail in section IV.

II. Project Goals/Contributions

A. The usefulness of statistical testing in ML

We will discuss the importance of statistical testing in the context of ML and DL. The unprecedented growth of ML and DL research/utilization has been unsupported by the absence of empirical statistical hypothesis testing and relies on single score comparison methods. Due to highly non convex loss surfaces (Li et al., 2017) *[4]* and local minimums under this curve, we require a selection method that undergoes more scrutiny.

B. Evaluation of ASO Testing

ASO is a statistical test that evaluates the cumulative distributions of our scores from two model architectures. Since the results of these models are highly dependent on the many hyperparameters it requires (layers, epochs, nodes, scale, etc.), we cannot infer an objectively 'better' model based on the mean scores of the models like when using a t-test or any singular statistic Dror et al. (2019) *[5]*. Instead, we use a violation ratio of the CDFs of our two score distributions to decide whether to reject the null hypothesis which will be discussed in section IV.

C. Test Comparison Studies

First, we investigate the type I & II errors of ASO compared to other significance tests with respect to sample size. As an experimental comparison of these tests, we sample from known distributions including normal, normal mixture, Laplace, and Rayleigh. Evaluating the Type I and II errors will help us choose a preferred method of statistical testing under our scenario studies and provide us a reasonable threshold value for ASO.

D. Scenario Studies

We conduct two simulation studies that use the CIFAR-10 image classification dataset under a CNN architecture, and a Bay Area housing regression dataset under an MLP architecture. We chose to only run 100 iterations to calculate the Type I and II errors due to time and computational limitations, although we would have preferred to run 1000 iterations so our error rates would have 2 decimal places of precision. These scenario studies are to illustrate the usefulness of statistical testing under real datasets in an applicable environment.

III. Usefulness of Statistical Testing in ML

A. Motivation

Statistical testing allows us to test the reliability and validity of a model by examining the performance of the model on different datasets. It can help us evaluate the efficacy of different ML algorithms, compare the performance of different models, and identify the factors that affect model performance. Statistical testing provides a rigorous and objective way to assess the significance of differences between model outcomes, and it enables us to draw more robust conclusions about the quality of ML models.

B. NLP Research

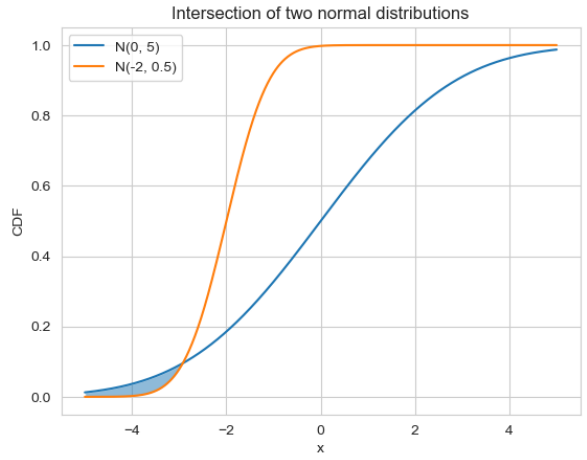
A research paper titled "An Empirical Investigation of Statistical Significance in NLP" Tberg (2012) *[6]* provides insights into the relationship between metric gain and statistical

significance in natural language processing (NLP). The study demonstrates the significance level's dependence on the systems being compared. The authors also propose a simple approach to approximate the response of these factors for tasks with a limited number of system outputs. The paper emphasizes that simple thresholds are not a substitute for statistical significance tests and recommends the use of such tests to validate metric gains in NLP. However, the study also highlights the limits of formal testing and cautions against relying solely on statistical significance in settings where previous systems are unavailable.

IV. Evaluation of Almost Stochastic Order (ASO) Testing

A. *ASO Violation Region*

ASO is a way to evaluate if a model is 'better' by using the distributions of the loss scores to calculate the violation region. This region is the area of overlap between the empirical CDFs of both models. We can illustrate this region in the figure below shaded in blue with the CDF curves of 2 normal distributions following the notation N(μ, σ^2).



Of course, in practice, we will not know the true distribution of our scores. We use the empirical CDF of our collected scores to evaluate this region. To perform this test in code, we utilize an existing library that calculates the violation ratio for us. More information on the package used can be found in CILR (2022) *[1]*.

B. *Stochastic Order vs. Almost Stochastic Order*

Stochastic order applies to the entire distribution of random variables. It describes the relationship between two random variables based on the probability that one variable is larger than the other for all possible values.

On the other hand, almost stochastic order is a more general concept that applies to a subset of the distribution of random variables. Specifically, it requires that the probability of the first variable being larger than the second variable is greater than or equal to a specified threshold for all values of the second variable. In other words, almost stochastic order allows for the intersection of empirical CDF's up to a certain degree.

C. *Statistical Testing*

Let's define a one-sided test statistic denoted as $\delta(S_A, S_B)$ such that S_A and S_B are score observations from model A and B. We formulate our null hypothesis under the assumption that model A is not better than B. (ICLR 2022) *[1]*

$$H_0: E[\delta(S_A, S_B)] \leq 0$$

In a standard statistical significance test the p-value is used to represent the probability of obtaining a test statistic as, or more extreme as the one observed assuming the null hypothesis is true. In the context of a model selection problem, we can represent a p-value in the following way.

$$P(\delta(S_A^*, S_B^*) \geq \delta(S_A, S_B) \,|H_0)$$

Where the * superscript denotes the replicated score observations. We can interpret the statement above that if the probability of our replicated test statistic $\delta(S_A^*, S_B^*)$ being greater than or equal to our observed test statistic is low (i.e. below 0.05) then we have significant evidence to say model A is better.

D. *ASO Test*

Similarly, the ASO test is a nonparametric method of testing used to compare the distributions of two score samples. The test is based on the concept of stochastic dominance, which is a partial ordering of probability distributions that captures the notion that one distribution is "better" than another in the sense that it has a higher probability of producing smaller error score outcomes. We will express the null hypothesis in terms of our violation ratio $V(F(S_A), F(S_B))$ and some threshold τ. Dror et al. (2019) *[3]*.

$$H_0: V\big(F(S_A), F(S_B)\big) \geq \tau$$
$$H_1: V\big(F(S_A), F(S_B)\big) < \tau$$

There are several ways to quantify the violation of stochastic dominance. Here we will focus on the approach by del Barrio et al. (2018b) which quantifies the violation ratio by

$$V\big(F(S_A), F(S_B)\big) = \frac{\int (F_{S_A}^{-1}(t) - F_{S_B}^{-1}(t))^2 \; dt}{W_2^2}$$

Where we integrate t over the violation set t ϵ (0,1) when

$$F^{-1}(S_A) \geq F^{-1}(S_B)$$

Our denominator W is the univariate l_2 Wasserstien distance

$$W_2 = \sqrt{\int_0^1 (F_{S_A}^{-1}(t) - F_{S_B}^{-1}(t))^2 dt}$$

For information on a minimal bound on $V\big(F(S_A), F(S_B)\big)$ as well as the estimated variance term $\hat{\sigma}_{n,m}^2$ given empirical CDFs from S_A and S_B, see ICLR (2022) *[1]*.

V. Test Comparison Studies

A. Comparison Study

To validate the reliability of the ASO test, we compare its type I and type II error rates across different sample sizes and distributions using established statistical significance tests such as Student's t-test, Bootstrap, Permutation, Wilcoxon, and Mann-Whitney U. The experiment involves sampling from four different distributions, namely Normal, Normal Mixture, Laplace, and Rayleigh, to obtain the error rates. Then, we calculate the p-values from each test for each sample size in [5, 10, 15, 20] then repeat iteratively to obtain the type I & II error scores.

Let us define a single simulation as the process that returns a p-value for each test and sample size after sampling from a single distribution. We then run 750 simulations for each sampling distribution to calculate the type error rates. We set the violation rate threshold for rejecting the ASO test at 0.2, while the p-value threshold for rejecting all other tests is 0.05. It is crucial to understand that the choice of threshold can affect the type I error rates. To demonstrate this, we present tables (1-6) showing the error rates for different thresholds, ranging from 0.05 to 0.5 in increments of 0.05. Based on these results, we can see that the violation area threshold of 0.2 for ASO is most comparable to the threshold of 0.05 used by the other tests justifying the threshold values used.

Figure (1) Samples from $N(0, 1.5^2)$

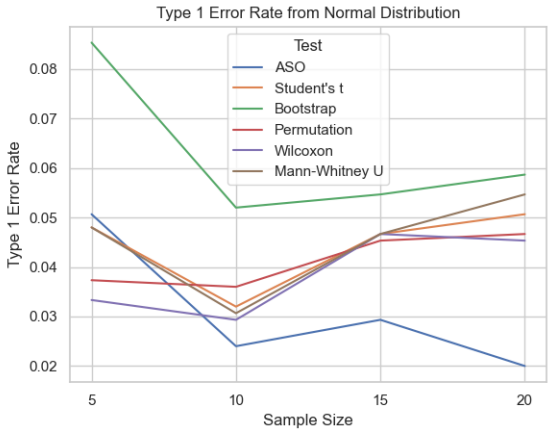


Figure (2) Samples from bimodal mixture model from $N(0, 1.5^2)$ *and* $N(-0.5, 0.25^2)$ *with mixture weights* $\pi_1 = 0.75$ *and* $\pi_2 = 0.25$.

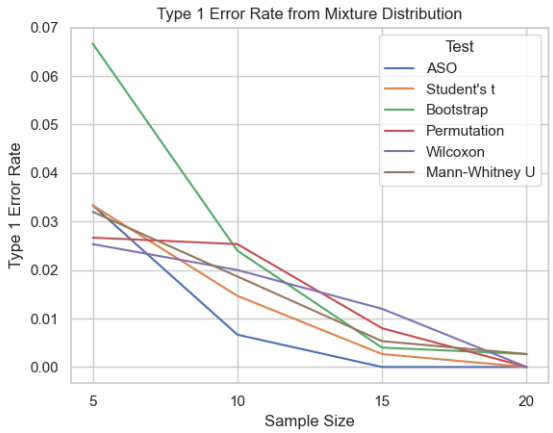


Figure (3) Samples from Rayleigh(1)

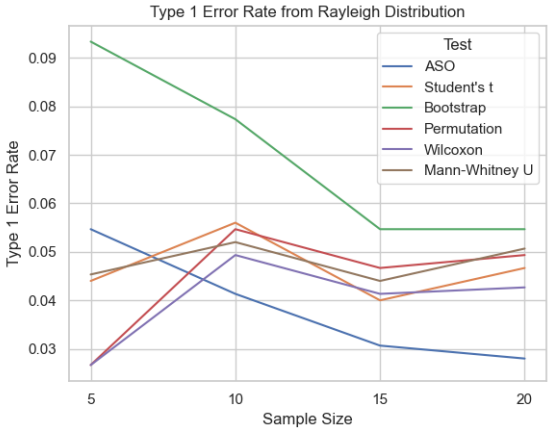
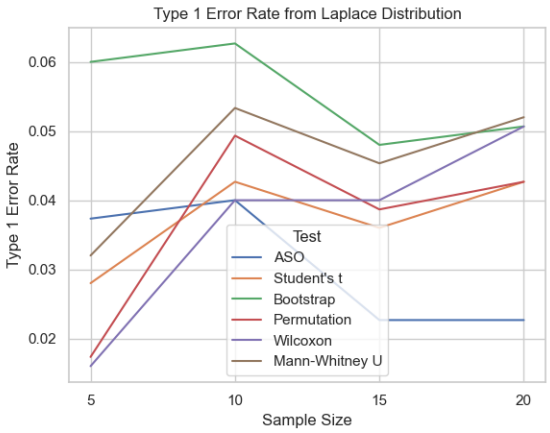


Figure (4) Samples from Laplace$(0, 1.5^2)$



B. Why Type 1 error matters

In the context of comparing two models, a type 1 error would occur when we incorrectly reject the null hypothesis that suggests that there is no difference between the models, i.e., we mistakenly conclude that model A is better than model B when in fact it is not. On the other hand, a type 2 error would occur

when we fail to reject the null hypothesis, i.e., we conclude that there is no difference between the models when in fact there is a difference.

In this scenario, a type 1 error is more critical than a type 2 error because it could lead us to choose the inferior model A over the better model B. This mistake could have serious consequences, especially in applications such as medical diagnosis or self-driving cars, where model accuracy is of utmost importance. This is why we focus on controlling the type 1 error rate before we start using ASO.

VI. CNN Model Selection

A. Dataset

To study CNN model selection using statistical significance testing, we experiment with the CIFIAR-10 dataset which is a popular image classification dataset that consists of 60,000 32x32 color images in 10 classes, with 6,000 images per class. The classes include common objects such as airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks.

B. CNN Architectures

We consider two CNN architectures: Model A and Model B. Model A has 6 convolutional layers with 32, 64, 128, 128, 256, 256 filters, followed by 4 fully connected layers. Model B has 4 convolutional layers with 32, 64, 128, and 128 filters, followed by 2 fully connected layers. Both models use ReLU activation function and max-pooling after each convolutional layer. For type 1 errors, we run the simulations on two Model A CNNs. For type 2 errors we run the simulations on Model A and Model B.

We train both models on the CIFAR-10 dataset using the same hyperparameters and the same training protocol. We use the ASO test to compare the performance of the two models on the test set. The ASO test is applied to the difference between the prediction scores of the two models on each test image. The ASO threshold is set at 0.2, and the p-value threshold for other tests is set at 0.05. After training on our population, we obtain a loss of 2.305 with Model A and 1.691 with Model B suggesting the simpler model B is better.

C. Process and Experiment Structure

To further validate the use of statistical tests, we calculate the type I and II errors once again. Even when delegating the training processes to a GPU running many simulations become computationally expensive. With this limitation, we can only capture 100 simulations for this study. It is important to note that in order to test two CNN models you only require running 1 simulation on a fixed number of bootstrapping iterations. In a real world setting we would run 100 or so bootstrap iterations to compare our test results with 2 significantly different model architectures. Running 100 simulations is only used to understand type I and II error behavior.

D. Simulations

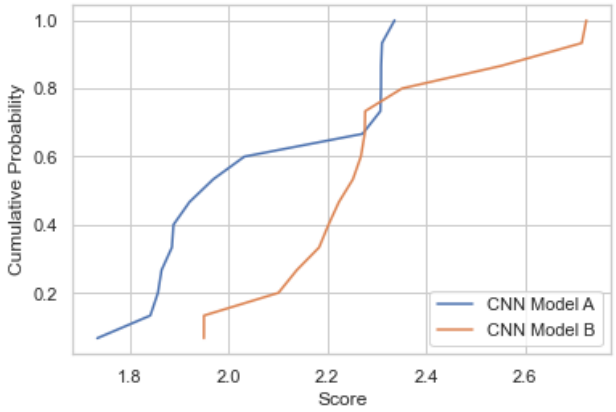
Each simulation involves running a set of bootstrap iterations [5, 10, 15, 20] on a new training subset of 2,000 pictures selected from the larger population of 60,000. The result is a set of loss scores of different lengths (depending on the number of bootstrap iterations) for each model. We then use these loss scores to run six different tests (including ASO) and obtain six p-values.

This process is repeated 100 times to calculate the type I and II error rates. In total, the experiment involves training the CNN models 4000 times (20 bootstraps x 100 simulations x 2 models). Each bootstrap iteration requires the model to be reinitialized, a bootstrap sample to be taken from the training subset, and the loss score to be calculated from the validation set.

E. Results

We obtain the type I and II error rates for the ASO test and other tests for different threshold values. Table 7 shows the error rates for threshold values up to 0.5 in increments of 0.05. As validation, we observe that the ASO test when calculating type I error rates is comparable at the chosen threshold of 0.2. Our type 2 errors seem to be very high. That is to say our two models are not significantly different enough under a sample size of 2000 to ensure Model A performs better than Model B. In fact, if we plot the empirical CDF of losses, we see that Model A is consistently performing better than Model B which is flipped from our initial inclination from training on our population. We believe this is due to the nature of our model architecture and how it reacts to a much smaller training sample.

Figure (5)



This notable shift in performance reflects the large type 2 error rates we see in Figure (7). This is because our one-sided test tries to investigate if Model B is better when seemingly Model A is under small sample sizes.

Figure (6)

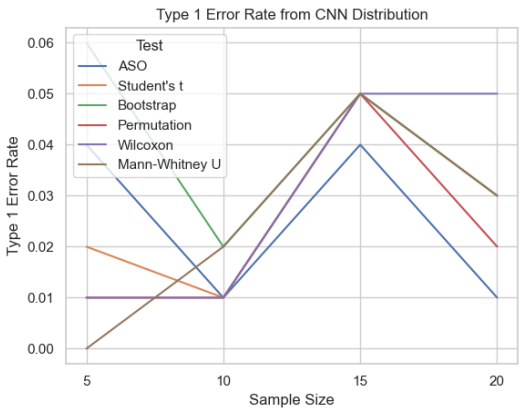
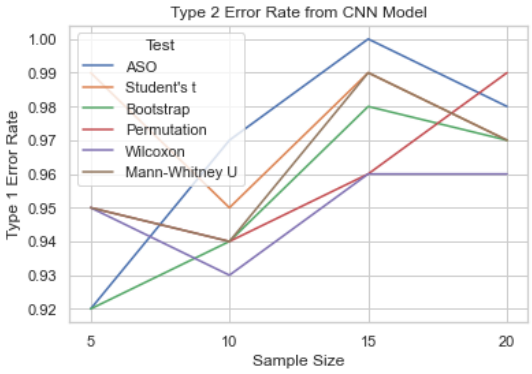


Figure (7)



F. Conclusion

In this case study, we demonstrate the use of statistical significance testing in selecting a model architecture for the image classification task on the CIFAR-10 dataset. We show that the ASO test can be used to compare the performance of two CNN architectures. These tests can be applied to other machine learning tasks and datasets to improve the model selection process. In the next section we explore the same concept under MLP model architectures.

VII. MLP Model Selection

A. Dataset

For our MLP model framework we utilize housing data from the Bay Area found on Kaggle. The dataset has 20640 rows and contains the columns longitude, latitude, housing_median_age, total_rooms, total_bedrooms, population, households, median_income, median_house_value, and ocean_proximity. The target variable we are trying to predict is the median_house_value. After one-hot-encoding and KNN imputation on missing data, we build the architectures of our 2 models.

B. MLP Architecture

We consider two MLP architectures: Model A and Model B. Model A has one hidden layer consisting of 5 neurons, and a maximum number of iterations during training set to 1000. Model B has one hidden layer consisting of 8 neurons, and the same maximum number of iterations during training.

Over our population, Model A has an MSE of 7625642793.58 while Model B has an MSE of 8007362787.87. Thus, for the purposes of the type II error simulation, Model A is considered to be the better model.

C. Experiment

The simulations, bootstrap iterations, and pipeline framework are essentially the same as in the CNN case study. We run 100 simulations to obtain type I and II error rates as a function of sample size and testing method. We first do an %80 train test split of the entire dataset and obtain our MSE under the test to verify performance differences due to model architecture.

D. Results

We obtain the type I and II error rates for the ASO test and other tests for different threshold values. Table 8 shows the type I error rates for threshold values up to 0.5 in increments of 0.05. As validation, we observe that the ASO test when calculating type I error rates is comparable at the chosen threshold of 0.2. Our type II error decreases rapidly as sample size grows showing significant difference between the two models. In other words, our significance tests correctly determine that the two models perform differently with increasing accuracy as sample size increases.

Figure (8)

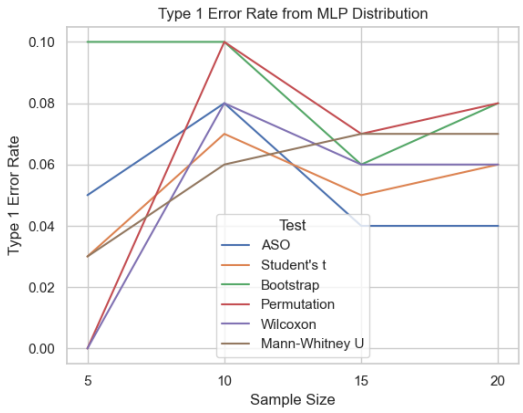
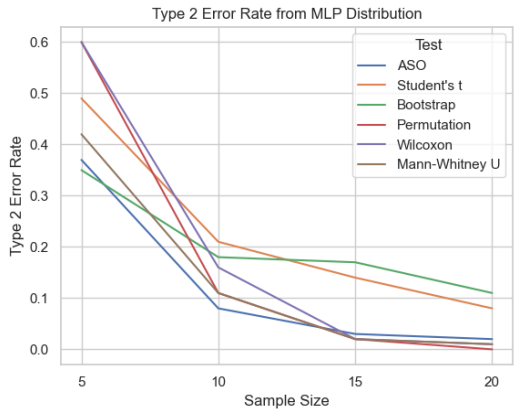


Figure (9)



E. Conclusion

Like the CNN study, the plots show that ASO with a threshold of 0.2 performs similarly to the other significance tests, in both type I and type II errors. For larger sample sizes, ASO actually outperforms the others in minimizing type I error, whereas it is middle-of-the-pack in minimizing type II error. For all tests, type II error seems to approach zero as the sample size is increased, meaning that our tests are able to correctly determine that model A is the better model by bootstrapping from a population sample.

VIII. Conclusion

Statistical significance testing in machine learning is an underused model comparison tool that can lead to more reliable model selection and strengthen ML research. We've demonstrated case studies from known distributions and applied tests to real datasets using convolutional neural networks and multilayer perceptron. Our results illustrate that the ASO test at a threshold of 0.2 is comparable to the established tests at 0.05 across the score distributions investigated in this paper. Although we shouldn't necessarily rely on statistical significance testing completely to derive conclusions about our models, it should certainly be utilized in conjunction with other methods like cross-validation, information criteria, holdout, etc. With more time, we would expand our scope of ML and DL models for a more comprehensive study. Our type I and II pipeline is quite computationally expensive mainly due to calculating the ASO test. Given more time, we would like to perform 1000 simulations instead of 100 to obtain more accurate results.

References

[1] D. Ulmer, C. Hardmeier, J. Frellsen "Easy and Meaningful Statistical Significance Testing in the Age of Neural Networks" URL https://arxiv.org/abs/2204.06815

[2] Nils Reimers and Iryna Gurevych. Why comparing single performance scores does not allow to draw conclusions about machine learning approaches. arXiv preprint arXiv:1803.09578, 2018.

[3] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1383–1392, 2018

[4] Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. The Journal of Machine Learning Research, 18(1):2653–2688, 2017.

[5] Rotem Dror, Segev Shlomov, and Roi Reichart. Deep dominance - how to properly compare deep neural models. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pp. 2773–2785. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1266. URL https://doi.org/10.18653/v1/p19-1266.

[6] Taylor Berg-Kirkpatrick, David Burkett, Dan Klein, An Emperical Investigation of Statistical Significance in NLP. URL https://aclanthology.org/D12-1091/

[7] Project Overview Video URL: https://www.youtube.com/watch?v=5Sb3s7QCiV4

[8] Github code files URL: https://github.com/sbslaton/230-254-Final-Projects/tree/main/254