Final Project

Adversarial Attacks on Neural Automatic Essay Scoring Systems

1. Lee, Michael
2. Nickerson, Micah

CSCI S-89a Deep Learning, Summer 2019

Harvard University Extension School

Prof. Zoran B. Djordjević

Adversarial Techniques

Using a specifically chosen model input to influence the model to output a desired result -- *i.e. Tricking a system into producing an artificially high score.*

Examples of Adversarial Techniques for Automated Essay Scoring Systems:

- Shuffling
- Appending
- Insertion
- Progressive Overload
- Single Substitution

Neural Automated Essay Scoring Systems

Supervised Learning Predictive Neural Network to "read" a specific essay set and classify a score on an assigned scale.

Implemented for standardized testing; Intended to lessen burden on human readers, by pre-scoring essays, and flagging those requiring manual verification.

Method of Procedure

1

Develop an Automated Essay Scoring System - "Black Box"

- **Construct** Neural Network Scoring System for Essays - per Lang et al. (2019)
- **Train** on Student Essay Dataset - "ASAP Data"
- **Evaluate** AES performance on test

2

Test Adversarial Techniques

- **Define** Adversarial Techniques
- **Generate** Adversarial Essays
- **"Attack"** AES System

3

Evaluate

- **Evaluate** AES performance against adversarial attacks
- **Recommend** improvements

Dataset - ASAP Student Essays - 10th Grade

ASAP
Automated Student Assessment Prize
Phase One: Automated Essay Scoring

The Hewlett Foundation: Automated Essay Scoring

$100,000 · 154 teams · 7 years ago

Essay Set #2

Type of essay:	Persuasive/ Narrative/Expository
Grade level:	10
Training set size:	1,800 essays
Final evaluation set size:	600 essays
Average length of essays:	400 words
Scoring:	R1D1 (Reader 1 Domain 1 Score), R1D2 (Reader 1 Domain 2 Score), R2D1 (Reader 2 Domain 1 Score), R2D2 (Reader 2 Domain 2 Score), D1_Resolved, D2_Resolved
Domain 1 (Writing Applications) Rubric range:	1-6
Domain 2 (Language Conventions) Rubric range:	1-4
Domain 1 (Writing Applications) Final score range:	1-6
Domain 2 (Language Conventions) Final score range:	1-4

2400 Total Essays

NOTE: This data set is the only one that is scored using a trait rubric. You will be asked to make two separate predictions for this essay prompt corresponding to the resolved scores for the two domains that were assessed.

Prompt

Censorship in the Libraries

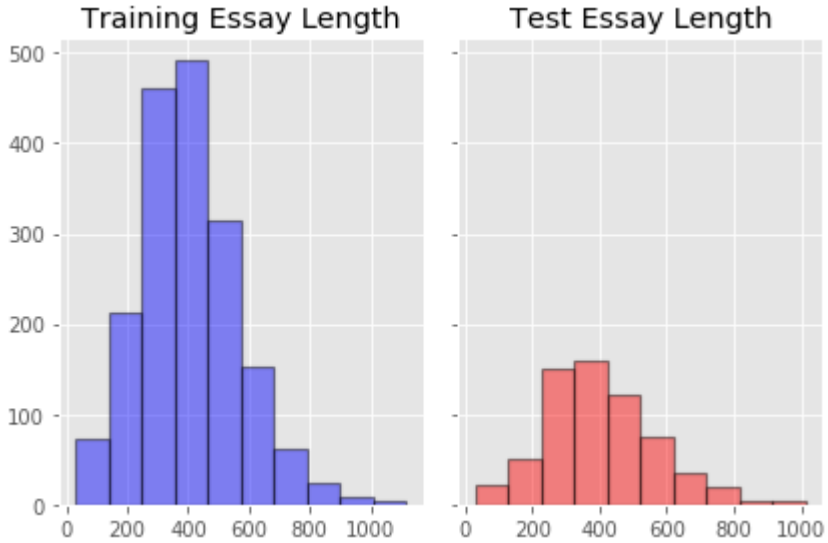
"All of us can think of a book that we hope none of our children or any other children have taken off the shelf. But if I have the right to remove that book from the shelf -- that work I abhor -- then you also have exactly the same right and so does everyone else. And then we have no books left on the shelf for any of us." --Katherine Paterson, Author

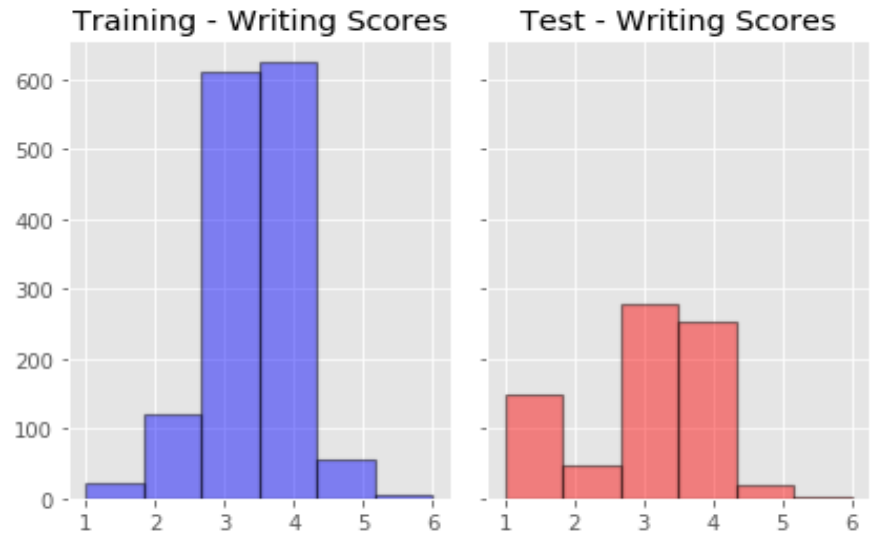
Write a persuasive essay to a newspaper reflecting your vies on censorship in libraries. Do you believe that certain materials, such as books, music, movies, magazines, etc., should be removed from the shelves if they are found offensive? Support your position with convincing arguments from your own experience, observations, and/or reading.

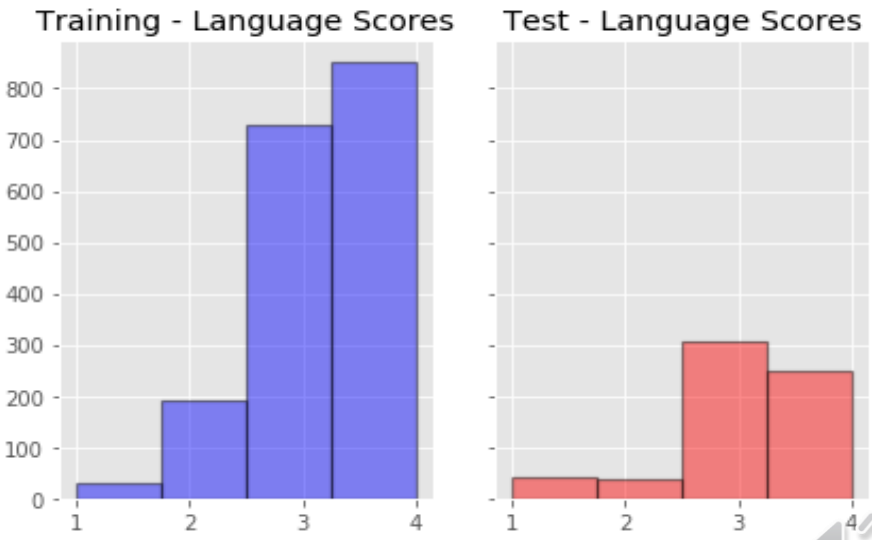
Essay Set - Exploratory Data Analysis

Essay Word Range: 30 - 1118 words

Average Essay is 402.1 words





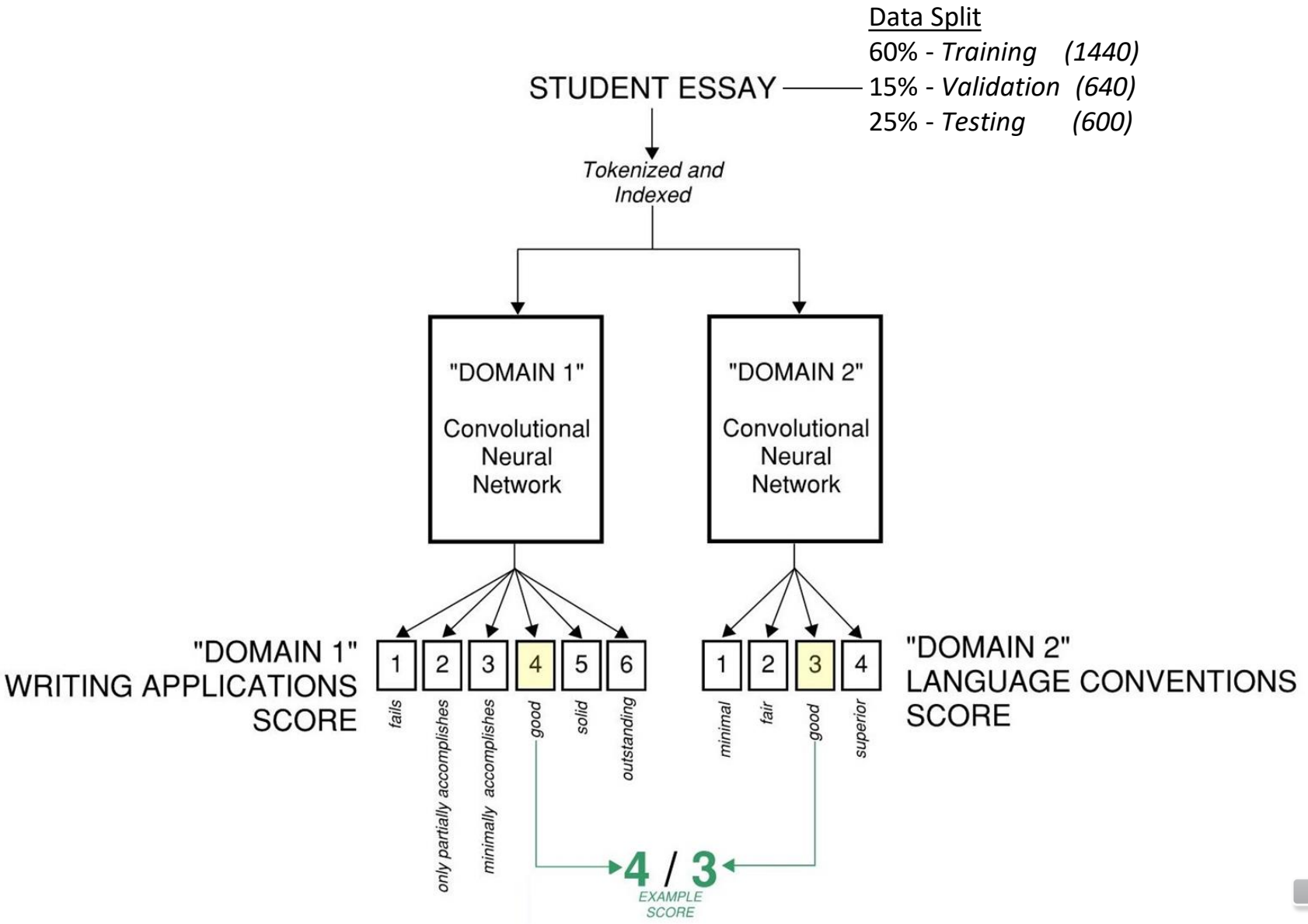


1

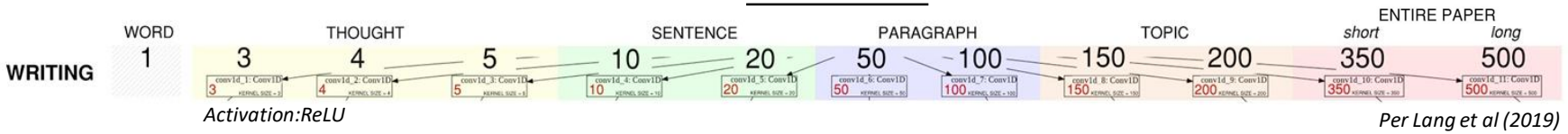
Develop a “Black Box”

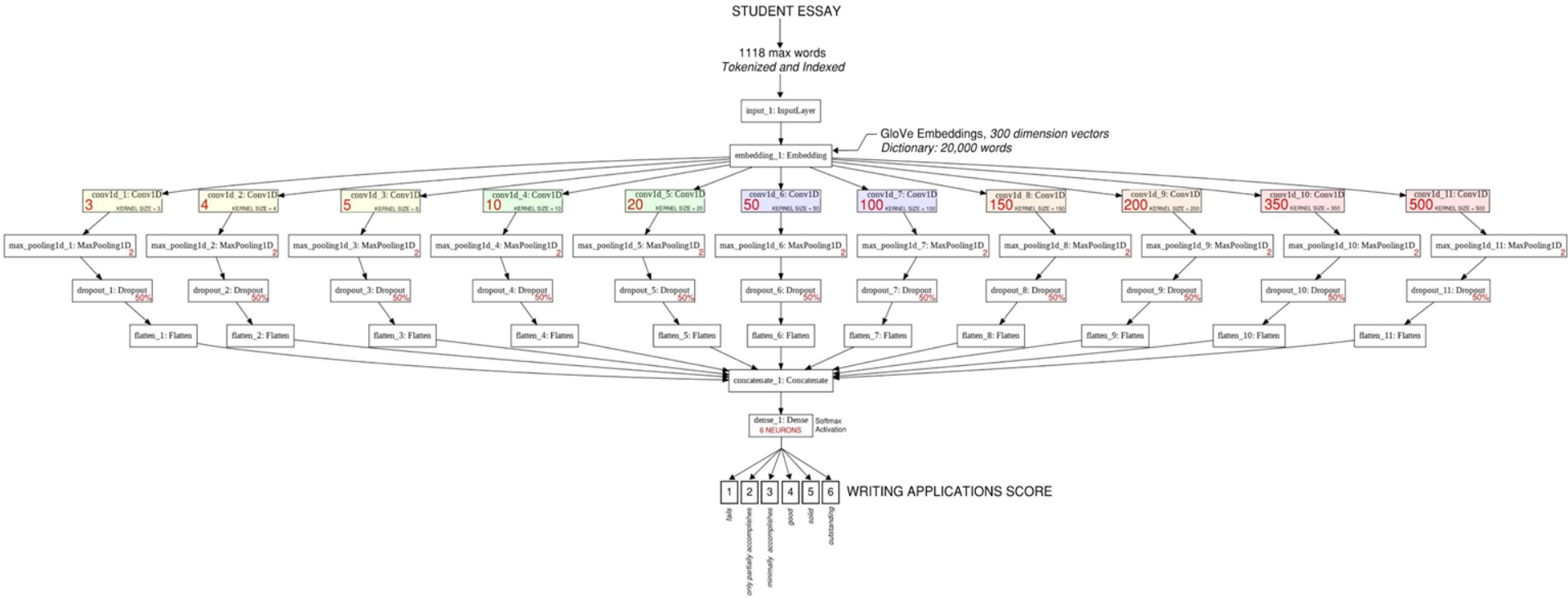
Automated Essay Scoring System

Automated Essay Scoring (AES) System

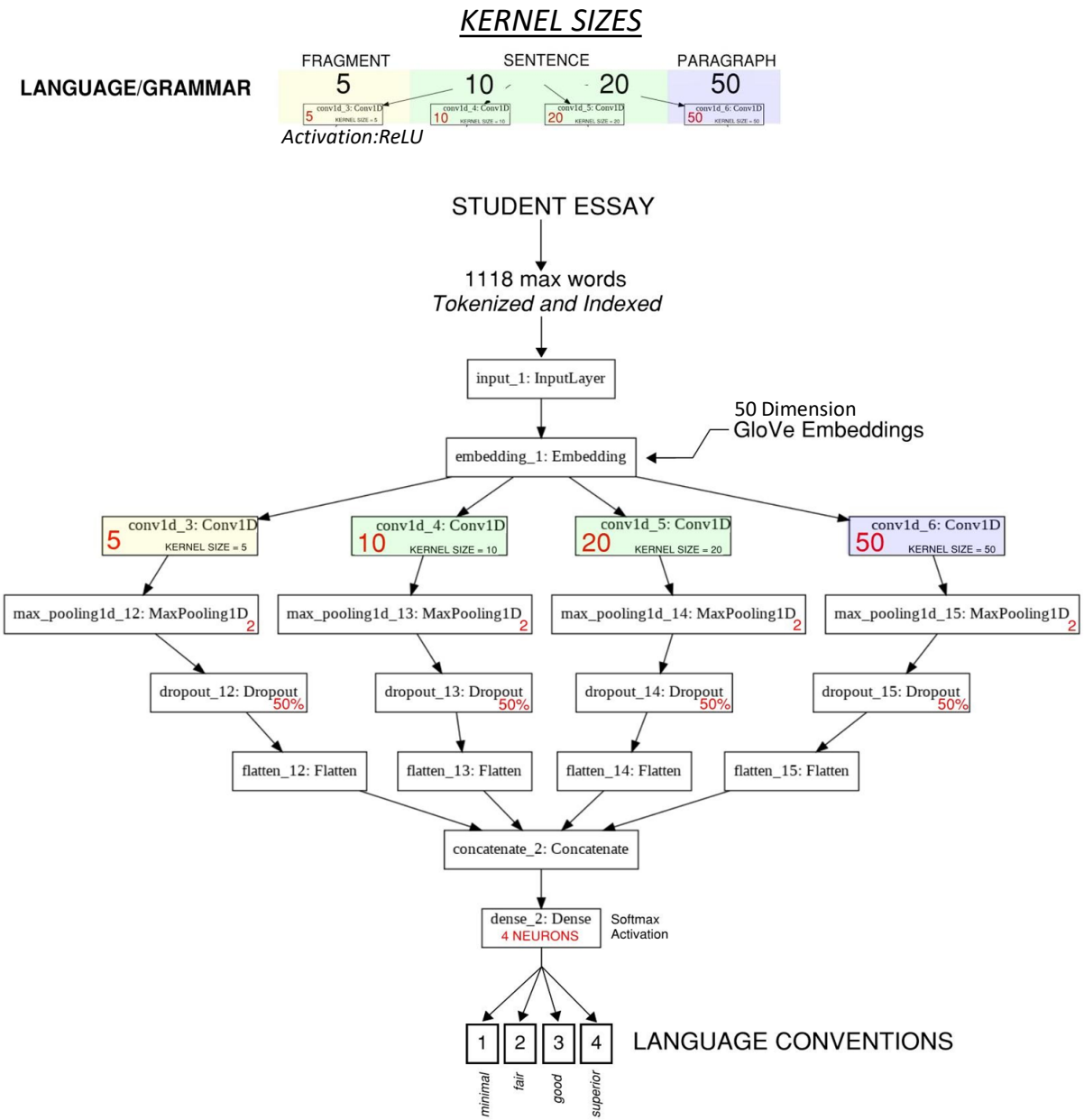


“Domain 1” Convolutional Neural Network

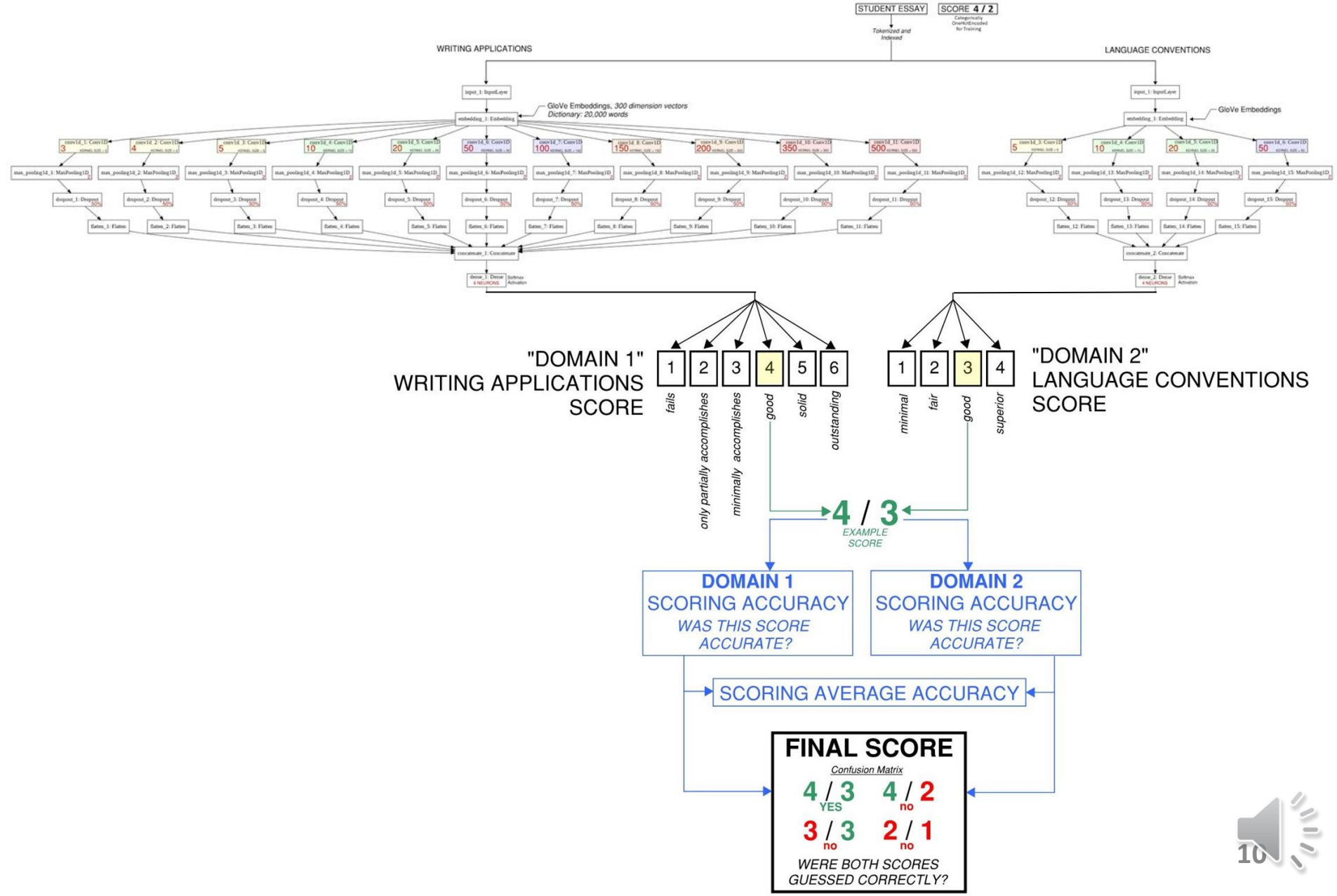




“Domain 2” Convolutional Neural Network



Essay Scoring & Model Evaluation



AES Blackbox Performance

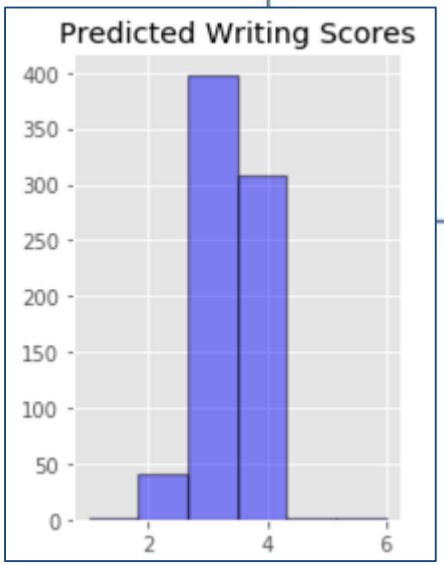
"Writing Content & Style"

DOMAIN 1
SCORING ACCURACY
76.83%

"Language Conventions & Grammar"

DOMAIN 2
SCORING ACCURACY
69.67%

AVERAGE ACCURACY *73.25%*

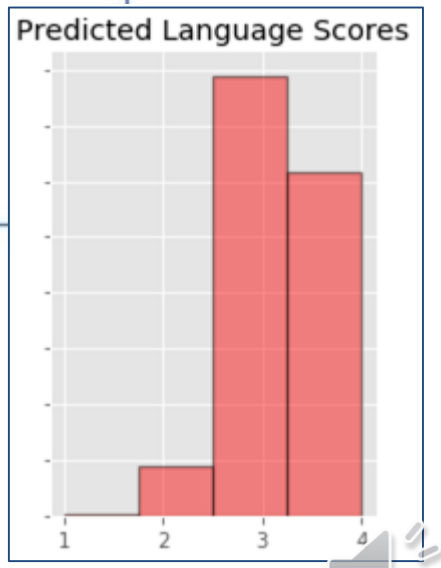


FINAL SCORE

Confusion Matrix

	d2	d2
d1	56.83% YES	12.83% no
d1	20.00% no	10.33% no

56.83%



Benchmark Scores

.5683 *vs*

Competition

ASAP
Automated Student Assessment Prize
Phase One: Automated Essay Scoring

The Hewlett Foundation: Automated Essay Scoring

Develop an automated scoring algorithm for student-written essays.

$100,000 · 154 teams · 7 years ago

#	△pub	Team Name	Kernel	Team Members	Score	Entries	Last
1	▲1	SirGuessalot & PlanetThanet ...			0.81407	148	7y
2	▲2	@ORGANIZATION			0.80881	102	7y

Professional System

"... exact agreement between human and machine scores **between 0.50 − 0.52**..."

ETS

Automated Essay Scoring With E-rater® v.2.0

System	ELW	N	Mean kappa	STD kappa	Exact agreement
GMAT argument					
Specific	.2	7	.38	.06	.52
	.3	7	.38	.07	.52
Generic10	.2	7	.35	.08	.50
	.3	7	.35	.08	.50
Generic12	.2	7	.38	.06	.52
	.3	7	.39	.07	.52
v.1.3	-	7	.36	.07	.51

Per Attali et al. (2005)

Random Guess

Chance of Randomly Guessing: ***0.0004***

Chance of Randomly Guessing - from only (3,4): ***0.25***

2 Test Adversarial Techniques

Adversarial Techniques

Anchors

- **Words representing the minimum input to model needed to achieve a desired model prediction score**

- Can be identified by substituting words in essay with UNK tokens
  - Not computationally efficient

- Chose anchors which were relevant to the essay prompt:
  - “Library”
  - “Censorship”

- Included a non-anchor word as a control:
  - “The”

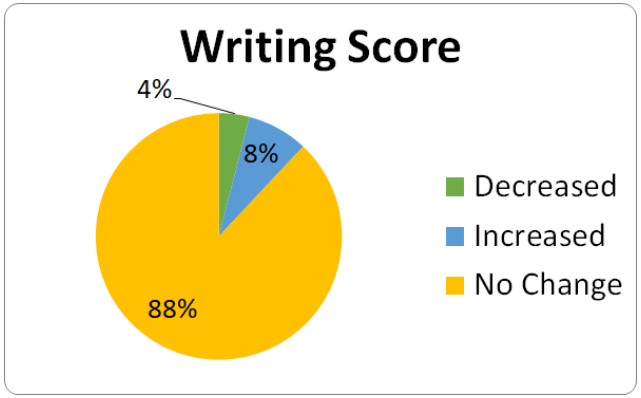
Adversarial Techniques

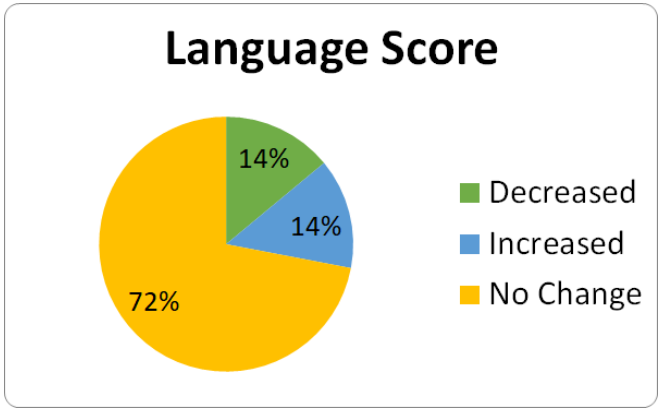
Shuffling Attack

- **Words in host essay are randomly shuffled**

Example Essay:

censorship explore. like that's their well Some for to solve that means people reading libraries. brain their not want because or time. shelf would should @MONTH1 taken the like that read. of off @CAPS1 them age desiding public those carry books not, catagorized I them. shelves, reason. Public fun other are books cattegorized locations restricted a lot and ideas that would from Books the when ignorant for sealing concerned





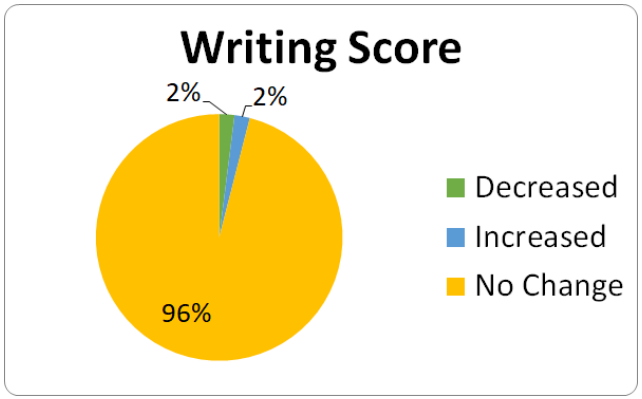
Adversarial Techniques

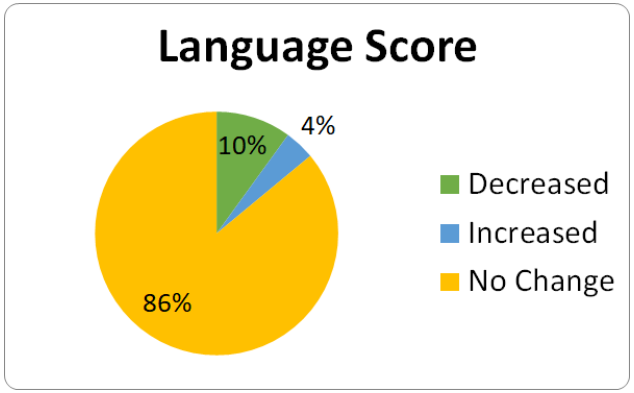
Appending Attack

- **Anchor word is appended to the end of the host essay.**

Example Essay:

Taking a book off of the shelf is like and author coming up and destroying all of the hard work you have worked for. Many people might find a book that is very inappropriate and want it to be disposed of but as you can see, its not right. There would be no books left on the shelves. It would be very disrespectful to destroy an authors hard work. People have a choice in what they read and we should not take that away **library**





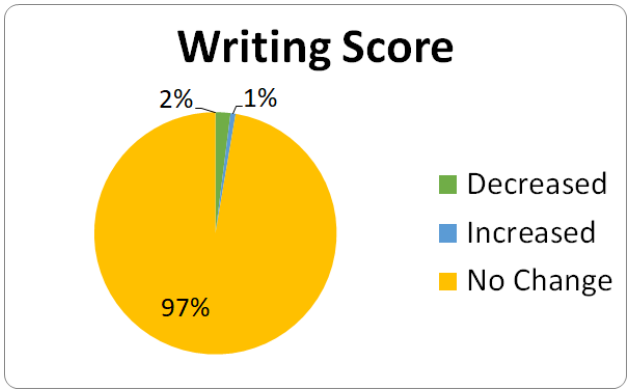
Adversarial Techniques

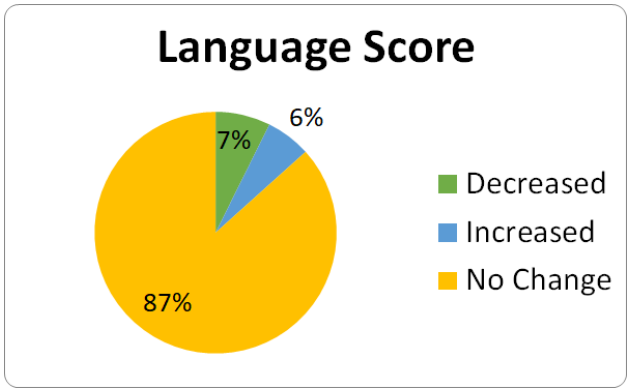
Insertion Attack

- **Insert anchor word into an host essay at a random location.**

Example Essay:

Books, music, movies, magazines, etc., should not be offensive. Kids should not **library** read a book, magazine, etc., if the author is talking about harming someone or something. If the book, magazine, etc., has inappropriate language or insult someone, that book needs to be taken off of the shelf and immediately thrown in the trash. The last thing a parent wants is their child to learn something bad from a book and repeat it at school, because that will only lead up to the child getting in serious trouble.





Adversarial Techniques

Progressive Overload Attack

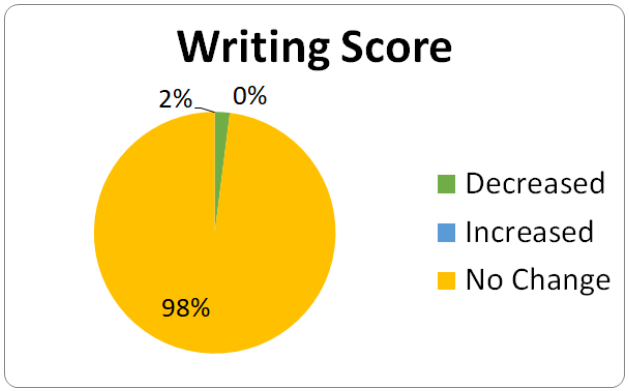
- **Words in host essay set are progressively replaced with the target anchor word.**
  - **First essay: first word** replaced with the anchor
  - **Second essay: first two words** replaced with anchor
  - **Third essay: first three words** replaced with anchor

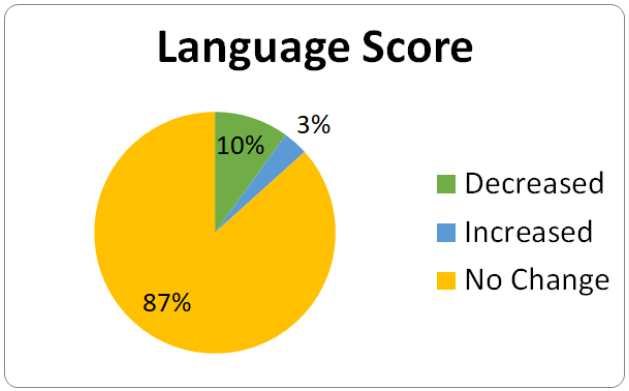
Example:

library u believe there are books, music, magizines, and movies in are libaries?

library library believe that there are some materials that should not be allowed on the shevles.

library library library that certain books, movies, magazines, etc., should be removed from shelves in a library if found offensive?





Adversarial Techniques

Single Substitution Attack

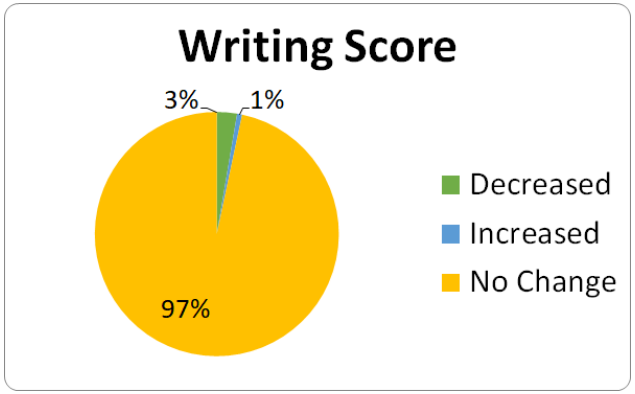
- **Words in host essay set are substituted for an anchor one essay at a time**
  - **First essay: first word replaced by an anchor**
  - **Second essay: second word replaced by an anchor**
  - **Third essay: third word replaced by an anchor**

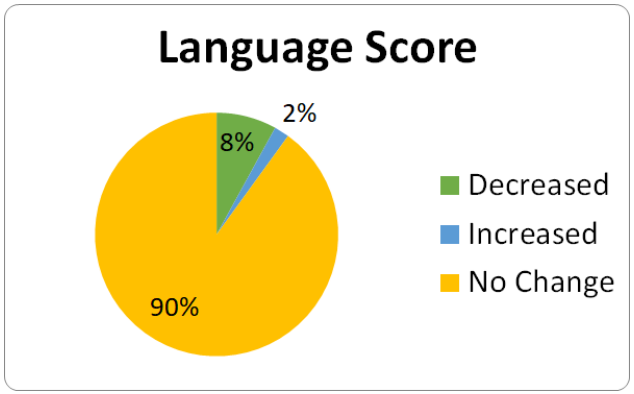
Example:

library u believe there are books, music, magizines, and movies in are libaries?

I **library** believe that there are some materials that should not be allowed on the shevles.

Do you **library** that certain books, movies, magazines, etc., should be removed from shelves in a library if found offensive?





3 Evaluate Results

Adversarial Techniques - **Performance**

Mean change in model score prediction after implementing adversarial attack:

Domain 1 - **Writing Score**		
Attack	**Anchor**	**Non-Anchor**
Shuffling	4%	
Appending	0%	0%
Insertion	**-2%**	0%
Progressive Overload	**-2%**	**-2%**
Single Substitution	**-2%**	**-2%**

Domain 2 - **Language Score**		
Attack	**Anchor**	**Non-Anchor**
Shuffling	0%	
Appending	**-6%**	**-6%**
Insertion	0%	**-4%**
Progressive Overload	**-8%**	**-4%**
Single Substitution	**-7%**	**-4%**

- **On average, attacks resulted in a lower essay score prediction**
- Impact on Language Convention score higher than Writing Content score
- Words we thought would be good candidate anchors turned out to result in lower Language Convention score than the control word.

Summary

Summary of Project

- Dual Score CNN for Automated Scoring performed better on writing than on grammar. It evaluated essays better benchmark professional evaluation but requires improvement to match current research examples.

- Adversarial attacks appeared to have a slightly negative impact on AES score predictions.

- **Demonstrates that the use of adversarial attacks was not effective in achieving a higher AES score prediction.**

Recommendations for Improvement

- A GRU based RNN model was tested but was poorly performing *(43%)* - further research to optimize should produce a more successful AES system than was built for project.

- Testing on larger sample set

References

Attali, J.B. a. Y. a., Automated essay scoring with e-rater v. 2.0. ETS Research. *ETS Research Report Series* **2004(2):i–21**, (2005).

Briscoe, Y. F. a. Y. a., Neural Automated Essay Scoring and Coherence Modeling for Adversarially Crafted Input. *CoRR* **abs/1804.06898**, (2018).

Gilmer, T. B. B. a. M. a. R. a. A. a., Adversarial Patch. *CoRR* **abs/1712.09665**, (2017).

Guestrin, S. S. a. M. T. a, Anchors: High precision model-agnostic explanations. *Thirty-Second AAI Conference on Artificial Intelligence*, (2018).

Dery, N.H. a. D. a., Neural Networks for Automated Essay Grading. *Stanford University* **cs224d/huyenn**, (2016).

Kim, Yoon. Convolutional neural networks for sentence classification. *arXiv preprint* **arXiv:1408.5882**, (2014).

Lang, K. K. a., Adversarial Examples for Neural Automatic Essay Scoring Systems. *Stanford University* **cs224n/15720509**, (2019).

Xiong, Z. W. a. W. a., Automated Essay Scoring System. *University of Illinois at Urbana-Champaign* **CS410/aess-project**, (2018).

Zhao, J. L. a. X. a., Automated Essay Scoring based on Two-Stage Learning. *CoRR* **abs/1901.07744**, (2019).

Project Index

- **Adversarial Attacks on Automated Essay Scoring Systems**

1. Lee, Michael
2. Nickerson, Micah

- **Two-minute (short) video**:
  - https://youtu.be/V0C0RNsNceY
- **Reference Links:**
  - Dataset : "The Hewlett Foundation: Automated Essay Scoring"
    - https://www.kaggle.com/c/asap-aes
  - Project Code (Github):
    - https://github.com/mjnickerson/csci-89a-final_project

CSCI S-89a Deep Learning, Summer 2019

Harvard University Extension School

Prof. Zoran B. Djordjević