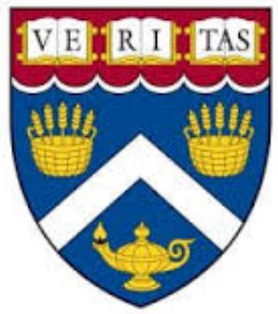
Final Project
Automated Essay Scoring

Gupta, Amit



CSCI S-89a Deep Learning, Summer 2019
Harvard University Extension School
Prof. Zoran B. Djordjević

Automated Essay Scoring

- **Hewlett Foundation released an essay dataset on Kaggle for an automated essay scoring competition in 2012. (first place prize of $100K)**
- **Generated significant interest**
- **Papers quote benchmarks against this dataset**

Shallow Networks (SVM, RF) (Feature Engineered)

- **State of art: 66% accuracy/ agreement with human rater**

Deep Learning Based (Dec 2018)

- **State of art: 80% accuracy/ agreement with human rater**

Project Goal - Get our accuracy numbers closer to state of art DL models

Essay Dataset (8 prompts)

Prompt	#Essays	Scores	Type
1	1783	2-12	Narrative
2	1800	1-6	Narrative
3	1726	0-3	Source Dependent
4	1772	0-3	Source Dependent
5	1805	0-4	Source Dependent
6	1800	0-4	Source Dependent
7	1569	0-30	Narrative
8	723	0-60	Narrative

Final Project focus in on prompt 3.
Based on benchmarks this is one of the hardest prompts.

Prompt 3 has unbalanced scoring samples
39 samples with “0” score
607 samples with “1” score
657 samples with “2” score
423 samples with “3” score

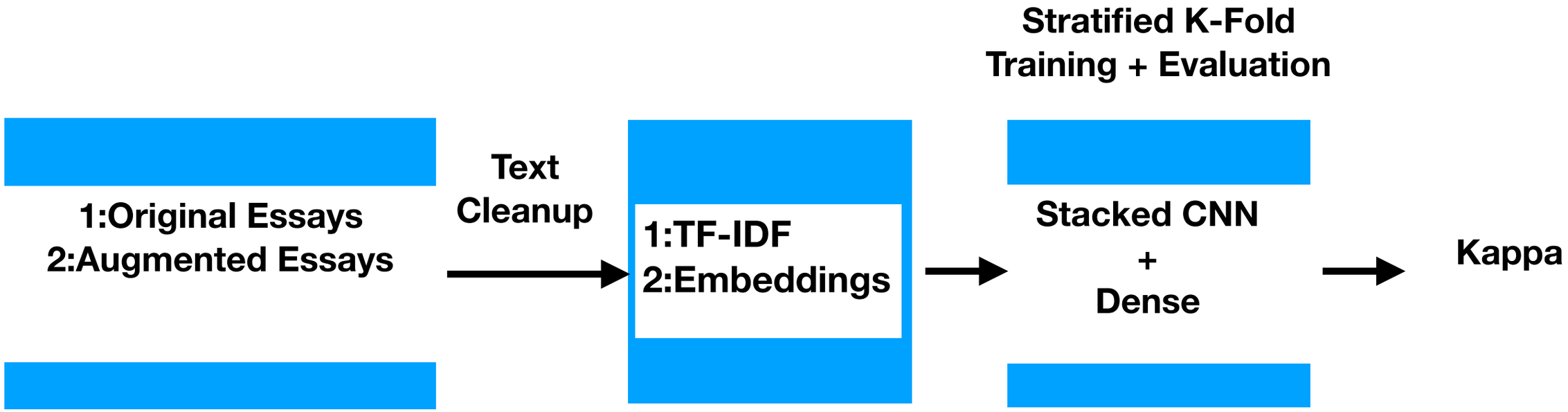
Addressing class imbalance with data augmentation

To address class imbalance of few zero score essays, following ideas (from Jason Wei and Kai Zou) on Synonym Replacement, Random Insertion, Random Swap and Random Deletion have been applied

Operation	Sentence
None	A sad, superior human comedy played out on the back roads of life.
SR	A ***lamentable***, superior human comedy played out on the ***backward*** road of life.
RI	A sad, superior human comedy played out on ***funniness*** the back roads of life.
RS	A sad, superior human comedy played out on ***roads*** back ***the*** of life.
RD	A sad, superior human out on the roads of life.

Augmentation operations for NLP proposed in **[this paper]**. SR=synonym replacement, RI=random insertion, RS=random swap, RD=random deletion. The Github repository for these techniques can be found **[here]**.

Dataflow Pipeline



Augmented Essays - Sample size was increased 10 fold from 39 samples to 390 zero scored samples using techniques described in the previous slide.

Stratified K-fold was used for model training and evaluation. This technique balances the class distribution across classes.

Kappa - This is a statistic that measures inter-annotator or inter-rater agreement. In this case, it is measuring the agreement between the model and the human rater

Results

	CNN+Dense (TF-IDF)	CNN+Dense (jointly learned embeddings)	State of art DL model	State of art feature-engineered Shallow networks
Kappa Score (original dataset)	**0.48**	**0.51**	**0.80**	**0.66**
Kappa Score (data augmented)	**0.76**	**0.76**		

We see a huge boost in performance going from 0.48-0.51 to 0.76 with data augmentation.

Kappa score of 0.76 beats state of art feature engineered model by wide margin and comes close to state of art DL model. This shows stacked CNN+Dense model is able to learn and discriminate key essay features, and compete with
the best DL models given enough training samples.

Results

Confusion Matrix: Original Dataset

	Score 0	Score 1	Score 2	Score 3
Score 0	1	7	0	0
Score 1	3	78	33	7
Score 2	1	77	101	12
Score 3	0	7	64	14

Out of 8 (1st row) zero score essays, 7 are mis-predicted with a “1” score

Confusion Matrix: Augmented Dataset (Significantly improved results)

	Score 0	Score 1	Score 2	Score 3
Score 0	75	3	0	0
Score 1	3	68	46	4
Score 2	0	11	102	18
Score 3	0	5	56	24

Out of 78 zero score essays, only 3 essays are mis-predicted with a “1” score

Key Links and Contact Info

Youtube URL:
2 min: https://youtu.be/53hlsm-CWfA

Dataset and Data Augmentation
AES Dataset:https://www.kaggle.com/c/asap-aes/data
NLP Data Augmentation: https://arxiv.org/abs/1901.11196

Key Papers
Neural Coherence : https://arxiv.org/abs/1711.04981
Siamese BiLSTM: https://www.mdpi.com/2073-8994/10/12/682

My contact: Amit Gupta
amitrgupta27@gmail.com

Appendix (Example Essay)

Type of essay:	Source Dependent Responses
Grade level:	10
Training set size:	1,726 essays
Final evaluation set size:	575 essays
Average length of essays:	150 words
Scoring:	1st Reader Score, 2nd Reader Score, Resolved CR Score
Rubric range:	0-3
Resolved CR score range:	0-3

Essay Set #3

Source Essay

ROUGH ROAD AHEAD: Do Not Exceed Posted Speed Limit
by Joe Kurmaskie

FORGET THAT OLD SAYING ABOUT NEVER taking candy from strangers. No, a better piece of advice for the solo cyclist would be, “Never accept travel advice from a collection of old-timers who haven’t left the confines of their porches since Carter was in office.” It’s not that a group of old guys doesn’t know the terrain. With age comes wisdom and all that, but the world is a fluid place. Things change.

At a reservoir campground outside of Lodi, California, I enjoyed the serenity of an early-summer evening and some lively conversation with these old codgers. What I shouldn’t have done was let them have a peek at my map. Like a foolish youth, the next morning I followed their advice and launched out at first light along a “shortcut” that was to slice away hours from my ride to Yosemite National Park……(cut here)

Prompt

Write a response that explains how the features of the setting affect the cyclist. In your response, include examples from the essay that support your conclusion.

Rubric Guidelines

Score 3: The response demonstrates an understanding of the complexities of the text.

- Addresses the demands of the question
- Uses expressed and implied information from the text
- Clarifies and extends understanding beyond the literal

Score 2: The response demonstrates a partial or literal understanding of the text.

- Addresses the demands of the question, although may not develop all parts equally
- Uses some expressed or implied information from the text to demonstrate understanding
- May not fully connect the support to a conclusion or assertion made about the text(s)

Score 1: The response shows evidence of a minimal understanding of the text.

- May show evidence that some meaning has been derived from the text
- May indicate a misreading of the text or the question
- May lack information or explanation to support an understanding of the text in relation to the question

Score 0: The response is completely irrelevant or incorrect, or there is no response.

Adjudication Rules

- If Reader-1 Score and Reader-2 Score are exact or adjacent, adjudication by a third reader is not required.
- If Reader-1 Score and Reader-2 Score are not adjacent or exact, then adjudication by a third reader is required.