Score Me if You Can: Study on Robustness of Automated Essay Scoring Systems to Out-of-domain and Adversarial Inputs

Vinit Hegiste
s8vihegi@stud.uni-saarland.de

Soumya Ranjan Sahoo
s8sosaho@stud.uni-saarland.de

Vladislav Skripniuk
s8vlskri@stud.uni-saarland.de

Abstract

Successes in natural language processing gave rise to numerous automated essay scoring systems, some of which are now used in high-stakes tests. In this project we take one of the recent models [20] into consideration and through several sanity checks reveal some of its intriguing properties.

1. Introduction

The process of evaluating students' writing is time-consuming and repetitive, thus it is appealing to shift this duty from lecturers and tutors to automatic essays assessors. Numerous systems were designed [13] [15] [20] [1] [4] [5] [22] and some of them are used in high-stakes tests [3]. It is therefore important to verify validity of grades assigned by such systems and make sure, that they are resistant to possible fraudulent actions. In this project we take one of the recent models [20] into consideration and through several sanity checks reveal some of its intriguing properties. In section 2 we make an overview of existing work on automated essay scoring, adversarial examples in text domain in general and for the task of automated essay scoring in particular. In section 3 we present our experimental results on robustness of automated essay scoring systems to out-of-domain data and adversarial examples. Section 4 summarizes conclusions we make in this project.

2. Related Work

The first AES system dates back to 1960s [13] when Project Essay Grade (PEG) was developed. Since then several systems were commercialized, one example is e-rater system [3], which is now used in Test of English as a Foreign Language (TOEFL) and Graduate Record Examination (GRE). Models like SVR were shown to assign reasonable marks based on handcrafted features [1] [15], while more recent work also leveraged neural models, like LSTM [20] [1], CNN [4] and more bizarre architectures [5] [22] for this task.

Involvement of AES systems into exam evaluation process raises an issue of accountability and validity of assigned marks. [16] tested earlier versions of e-rater system by revealing the inner structure of the system to experts and asking them to trick the system into assigning higher scores to their texts. [21] created a dataset of 30 adversarial essays for their SVM model. [7] show, that model of [20] can not distinguish random permutations of sentence from actual essays and introduce Local Coherence model to address this issue.

Generation of adversarial examples in text domain for tasks other than automated essay scoring attracted significant attention in the past years. [8] show possibility of circumventing Google's Perspective API by intentionally introducing typos. [17] attribute words to certain gender and substitute them to prevent classifier from inferring gender of the writer. [19] manipulate performance of sentiment analysis and gender detection systems with help of hand-crafted rules to substitute words in texts. [14] [6] use graph unfolding for RNNs and find corrections closest to gradient descent step in embedding space. [10] append adversarial sentences to paragraphs to confuse question answering system. [18] achieve the same goal by introducing semantic preserving changes to questions. [2] and [12] construct semantically close adversarial examples for the tasks of sentiment analysis and textual entailment. [9] construct adversarial examples with specified syntax with paraphrase networks trained with back-translation.

[1] https://github.com/edx/ease

3. Experiments

3.1. Data

The Automated Student Assessment Prize [2] was organized to facilitate research in the field of automated essays scoring. The dataset contains essays, written in response to one of 8 prompts by students in grades from 7 to 10. Each essay set was evaluated by human annotators using its own grading scale. Most existing models solve this task as a supervised learning task, by predicting mark rescaled to range from 0 to 1. To evaluate models predictions marks are scaled back to their original grading scale and Quadratic weighted Kappa score is applied:

$$\kappa = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}}$$

where $O_{i,j}$ is a number of essays, which received score i from human annotator and score j from model. $E_{i,j}$ is an element from an outer product of two histograms of scores assigned to essays by human annotator and by model (meaning an expected number of essays receiving score i and j from annotator and model respectively, if the model and annotator were totally uncorrelated, though marginal distribution of scores assigned by each of them was unchanged). Weights W_{ij} are computed as follows:

$$W_{ij} = \frac{(i-j)^2}{(N-1)^2}$$

where N is the maximal possible grade. This metric measures agreement in ratings, provided by two annotators with values ranging from -1 to 1, where 1 means perfect agreement, 0 means that two rankings are unrelated, and -1 means ratings of one annotator reverse the order induced by rankings of the other.

3.2. Model

For datasets from the field of computer vision pretrained models are usually made publicly available on the Internet [3] by authors or enthusiasts. However, for the task of automated essay scoring we were not able to find pretrained models, so the first step was to reproduce results of existing work. We decided to train LSTM model from [20]. To that end, we used code from two public repositories as reference implementations. [4] [5]

In the light of space limitations, we do not thoroughly describe model architecture and implementation details, we refer interested reader to [20] and our repository [6]. The model consists of LSTM cell, followed by a dropout layer, mean over time pooling layer and fully connected layer, which outputs predicted score on a scale from 0 to 1. The model is trained as a regressor using MSE objective.

3.3. Data preprocessing

During preprocessing phase we've used NLTK tokenizer to split essays into tokens and GloVe vectors to embed discrete tokens into $\mathbb{R}^{300}$. However, we noticed, that non-negligible share of all tokens can not be found in GloVe vectors dictionary. Inspection of several essays revealed existence of numerous spelling mistakes in some of them.

Extract from an essay without typos: *Everyone laughs differently , some people laugh so hard that they cry, and some don't. But one thing that they have in common is laughter.*

Extract from an essay with plenty of typos: *they will tall you wat the other peolepa are saind obut you and the other per. i haved a good frand and that trust me. trust is a good thik if you dont have aing trust you are noting.*

In ASAP-AES dataset description it is written, that students wrote these essays under strict time limitations and were not allowed any time for rereading or error correction, so possible misspellings were not taken into account during evaluation.

In table 1 we show proportion of unrecognized words in texts of all essays and in a vocabulary of all unique tokens, found in these essays. High percentage of unique unrecognized tokens in vocabulary is explained by Anna Karenina principle: there is only one way to be normal, while there are so many options to deviate from norm.

So, we decided to test two options: one is to substitute all out-of-vocabulary tokens with 'UNKNOWN' tag, another is to try to correct typos in all unknown words. To correct possible spelling mistake in unrecognized word we first found all closest words with edit distance not greater than 2, making typo in which could possibly result in this unrecognizable token. To choose one option out of pool of candidates we assigned each correction a probabilty of occurence, based on frequencies of all words in the dataset. We than picked substitution with highest probabilty. Another option to score candidate corrections based on context would be using Google Language Model [11], though we didn't invest much time into it, because subjective examination showed, that improper substitutions result from poor choice of candidate pool, not from poor scoring of candidates.

Using this approach we were able to substitute more than 99% of all unknown tokens by some known words, resulting in a dictionary of more than 10k substitutions. We provide first 20 consecutive entries from this dictionary below:

'somebad' → 'somebady', 'presentid' → 'presented', 'eithy' → 'eith', 'sorryndings' → 'surrondings', 'our-

[2] https://www.kaggle.com/c/asap-aes
[3] https://github.com/tensorflow/models/tree/master/research/slim
[4] https://github.com/nusnlp/nea
[5] https://github.com/zlliang/essaysense
[6] https://github.com/skripniuk/MLCySec

Essay set	Before		After	
	All	Unique	All	Unique
1	0.0048	0.1897	0.0002	0.0088
2	0.0050	0.1747	0.0001	0.0057
3	0.0069	0.1547	0.0016	0.0092
4	0.0055	0.1286	0.0002	0.0075
5	0.0040	0.1498	0.0002	0.0090
6	0.0040	0.1153	0.0001	0.0034
7	0.0066	0.1590	0.0003	0.0082
8	0.0012	0.0369	0.0001	0.0048

Table 1. Proportion of unknown tokens before and after correcting typos

Essay set	Before	After
1	0.7694 ±0.0217	**0.7702** ±0.0383
2	0.5651 ±0.0256	**0.5808** ±0.0316
3	**0.6544** ±0.0091	0.64567 ±0.0150
4	**0.6465** ±0.0148	0.6404 ±0.0195
5	0.7572 ±0.0262	**0.7588** ±0.0233
6	**0.7184** ±0.0281	0.7160 ±0.0238
7	0.6225 ±0.0312	**0.6232** ±0.0295
8	0.4247 ±0.0426	**0.4338** ±0.0525
Avg. QWK	0.6448 ±0.0137	**0.6461** ±0.0131

Table 2. QWK scores before and after correcting typos

tose' → 'purpose', 'colifornia' → 'california', 'resifes' → 'resipes', 'ceetain' → 'certain', 'edicational' → 'educational', 'documentories' → 'documentaries', 'continuence' → 'continuance', 'torirs' → 'toris', 'microcam' → 'microcar', "doens't" → "doesn't", 'steair' → 'stair', 'discuraged' → 'discouraged', 'characterice' → 'characterics', 'unasul' → 'unusual', 'stratend' → 'strated'

It can be seen, that some typos were fixed correctly, while some other substitutions are questionable. We than trained LSTM model on the original and corrected datasets using k-fold cross validation for evaluation. It can be seen in table 2, that essays with corrected typos are scored more accurately, though improvement is negligible. This result is a bit confusing, because during initial phase of the project, when we used only one fold and one essay set, we observed substantial improvement in QWK, which turned out to be non-significant when evaluated more properly. We also could not observe any differences in learning curves on figure 1, so we decided not to use typo corrections in later sections to retain compatibility with previous work.

3.4. Prompt specific models

The ASAP-AES is divided into 8 essay sets, essays from each essay set correponding to one of 8 prompts. Though each essay set is relatively small, on average consisting of 1.5k samples, we trained 8 LSTM models, one for each essay set. We report QWK scores on validation set in table 3.4. We achieved QWK values close to those of the authors, who also used a separate model for each essay set. These results show, that model trained on all essay sets does not profit from increased sample size, while separate models are doing quite well by learning prompt specific features. That raises a question of domain adaptation for Automated Essay Scoring Systems, which was discussed in [15].

3.5. Robustness on out-of-domain data

To investigate on existence of inputs, which may lead the model to produce unexpected predictions, we performed the following sanity check. We used the model to evaluate essay, consisting of multiple repetitions of one word. We executed this check for every word in a dictionary. Distribution of scores given by models on each of 8 essays sets if shown in Appendix.

Surprisingly, the marks spread over the whole range of possible grades, with significant number of essays receiving highest possible marks. At this point, it would be interesting to compare these distributions with confusion matrices computed on validation set of regular essays. For normal essays predicted score differs from the one provided by human rater by more than 1 point in very rare cases. Also distribution of scores for normal essays is much more concentrated in the middle, giving only a small share of students excellent marks. In contrary, distribution of scores on one-word essays is flat in the middle, with salient peaks at opposite ends of the range.

Indisputably, existence of such inputs limits applicability of this model in real-world scenarios. The form of aforementioned distribution of scores for adversarial essays makes us hypothesize, that learned models do not look for long term features like persuasive reasoning or coherent narrative, but are rather being triggered by certain words, which are divided into two big groups of those having positive and negative influence on the overall mark.

3.6. Adversarially perturbed essays

To investigate on existence of semantics preserving adversarial examples we have adopted method of [2]. This method substitutes words with their synonims, making use of genetic algorithm to select a set of best substitutions. Using this method we were able to achieve only a marginal improvement in scores. Some "successful" attacks can be found in Appendix. This bad luck does not mean that the adversary is generally weak. The method of [2] generates successful attacks for tasks of sentiment analisys or textual entailment, in these tasks model has to pay attention to details, because even single word "not" may change the meaning of the whole review by 180 degrees. Therefore, change

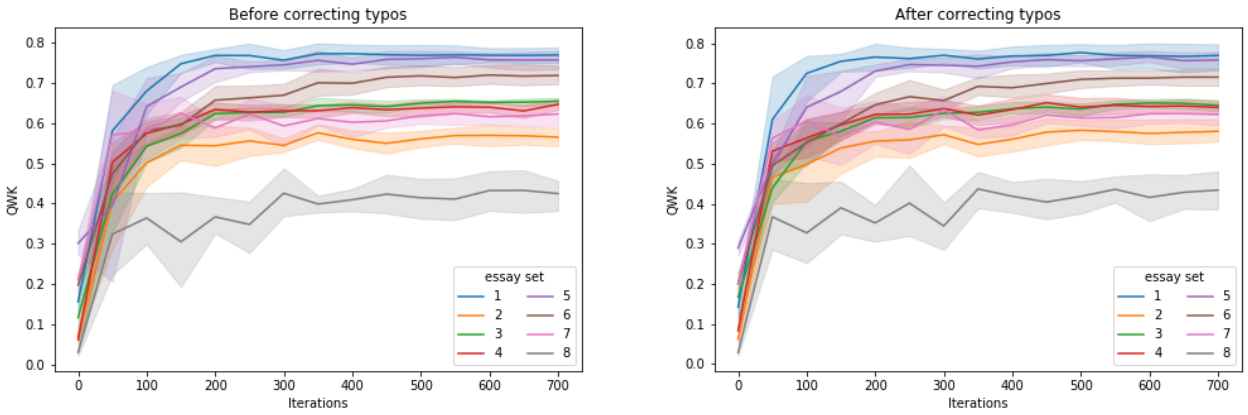

Figure 1. QWK of LSTM model on validation set

Method	Essay set								Avg. QWK
	1	2	3	4	5	6	7	8	
LSTM (our implementation)	0.769	0.565	0.654	0.646	0.757	0.718	0.623	0.425	0.645
8 LSTMs (our implementation)	**0.815**	**0.689**	0.635	**0.801**	0.801	**0.828**	0.761	**0.668**	**0.750**
LSTM (Taghipour & Ng)	0.775	0.687	**0.683**	0.795	**0.818**	0.813	**0.805**	0.594	0.746

Table 3. QWK scores on validation set. 8 LSTMs mean that a separate model is trained for each essay set.

of single words significantly affects model output. However, in the task of automated essay scoring we could not observe significant influence of single words on the overall mark, though it can be seen, that adversary is trying to drop fancy, pompous words here and there.

3.7. Generation of artificial essays

Let's consider this problem from a standpoint of lazy student, who wants to get a high mark without writing a word.

The described above approaches to get higher scores have certain drawbacks. One-word essays are too much perceptible, just a short glance is enough to notice that something is wrong with this essay. The method from section 3.6 only substitutes a small share of words with synonyms, so modified essay can be easily classified as plagiarism. To address these two issues we trained generative RNN on a subset of essays, which were evaluated as very good by human annotators (grades 5 and 6 in essay set 2). On generated essays we got a degenerate distribution of scores: we got grade 4 for 20 out of 20 essays generated with temperature in the fully connected layer of generative RNN equal to 1, and we got we got grade 5 for 20 out of 20 essays generated with temperature equal to 10. That result is counter-intuitive, because $T = 10$ makes distribution predicted by RNN almost uniform, so every next word is generated uniformly and independently of all previously generated words, but still that makes our LSTM scorer assign higher scores to such essays. Examples of generated essays can be seen in Appendix. LSTM scorer assigns higher score to essay, which is not at all coherent, but uses more extensive vocabulary, taken from essays, which received high grades from human annotators. We also generated uniform sequences of words using vocabulary of words from all essays (not only highly rated essays, but also low rated ones) and we got an average score of 2.

4. Conclusion

Conclusions, we make based on results of the experiments, are threefold:

1. Since performance of models trained separately on each essay set is significantly better than this of a model trained on all essays, we conclude, that task of essay evaluation is very prompt specific, i.e. model learns features specific for this topic and these feature do not transfer well between different topics.

2. The model we considered was easily triggered by one-word essays. That hints that rather than learning complex patterns, model makes it's predictions based on meaning of single words. Experiments with artificially generated essays also support this claim, since totally incoherent, but using extensive "right" vocabulary essays got higher scores than more coherent ones, which did not use that diverse vocabulary.

3. Essays are rather long texts compared to reviews or tweets, so the model does not concentrate on any particular words, making it difficult to manipulate predictions by changing words. In favor of this also speaks the fact, that correction of typos, which constituted a small share of all words in texts didn't influence performance of the model much.

References

[1] D. Alikaniotis, H. Yannakoudakis, and M. Rei. Automatic text scoring using neural networks. *CoRR*, abs/1606.04289, 2016. 1

[2] M. Alzantot, Y. Sharma, A. Elgohary, B. Ho, M. B. Srivastava, and K. Chang. Generating natural language adversarial examples. *CoRR*, abs/1804.07998, 2018. 1, 3

[3] Y. Attali and J. Burstein. Automated essay scoring with e-rater v.2.0. *Journal of Technology, Learning, and Assessment*, 2006. 1

[4] F. Dong and Y. Zhang. Automatic features for essay scoring - an empirical study. In *EMNLP*, 2016. 1

[5] F. Dong, Y. Zhang, and J. Yang. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162. Association for Computational Linguistics, 2017. 1

[6] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou. Hotflip: White-box adversarial examples for NLP. *CoRR*, abs/1712.06751, 2017. 1

[7] Y. Farag, H. Yannakoudakis, and T. Briscoe. Neural automated essay scoring and coherence modeling for adversarially crafted input. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 263–271. Association for Computational Linguistics, 2018. 1

[8] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran. Deceiving google's perspective API built for detecting toxic comments. *CoRR*, abs/1702.08138, 2017. 1

[9] M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. *CoRR*, abs/1804.06059, 2018. 1

[10] R. Jia and P. Liang. Adversarial examples for evaluating reading comprehension systems. *CoRR*, abs/1707.07328, 2017. 1

[11] R. Józefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu. Exploring the limits of language modeling. *CoRR*, abs/1602.02410, 2016. 2

[12] V. Kuleshov, S. Thakoor, T. Lau, and S. Ermon. Adversarial examples for natural language classification problems, 2018. 1

[13] E. B. Page. The use of the computer in analyzing student essays. *International Review of Education / Internationale Zeitschrift fr Erziehungswissenschaft / Revue Internationale de l'Education*, 14(2):210–225, 1968. 1

[14] N. Papernot, P. D. McDaniel, A. Swami, and R. E. Harang. Crafting adversarial input sequences for recurrent neural networks. *CoRR*, abs/1604.08275, 2016. 1

[15] P. Phandi, K. M. A. Chai, and H. T. Ng. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 431–439. Association for Computational Linguistics, 2015. 1, 3

[16] D. E. Powers, J. Burstein, M. Chodorow, M. E. Fowles, and K. Kukich. Stumping e-rater: challenging the validity of automated essay scoring. *Computers in Human Behavior*, 18:103–134, 2002. 1

[17] S. Reddy and K. Knight. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26. Association for Computational Linguistics, 2016. 1

[18] M. T. Ribeiro, S. Singh, and C. Guestrin. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865. Association for Computational Linguistics, 2018. 1

[19] S. Samanta and S. Mehta. Towards crafting text adversarial samples. *CoRR*, abs/1707.02812, 2017. 1

[20] K. Taghipour and H. T. Ng. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891. Association for Computational Linguistics, 2016. 1, 2

[21] H. Yannakoudakis, T. Briscoe, and B. Medlock. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 180–189, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. 1

[22] S. Zhao, Y. Zhang, X. Xiong, A. Botelho, and N. Heffernan. A memory-augmented neural model for automated grading. In *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*, L@S '17, pages 189–192, New York, NY, USA, 2017. ACM. 1

Appendix

Confusion matrices

Here we provide confusion matrices on validation set for each of 8 LSTM models, one for each essay set.

Actual marks	2	3	4	5	6	7	8	9	10	11	12
Predictions											
5	1	1	3	2	0	0	0	0	0	0	0
6	0	0	2	2	11	2	2	0	0	0	0
7	0	0	0	0	15	9	28	2	0	0	0
8	0	0	0	0	0	5	47	11	4	0	0
9	0	0	0	0	0	1	40	43	30	2	0
10	0	0	0	0	0	0	2	20	36	17	6
11	0	0	0	0	0	0	0	1	5	2	5

Figure 2. Essay set 1

Actual marks	1	2	3	4	5	6
Predictions						
2	17	76	19	0	0	0
3	1	49	445	143	0	0
4	0	2	140	479	35	1
5	0	0	0	10	20	3

Figure 3. Essay set 2

Actual marks	0	1	2	3
Predictions				
1	29	364	89	9
2	1	104	391	120
3	1	13	47	213

Figure 4. Essay set 3

Actual marks	0	1	2	3
Predictions				
0	140	50	0	0
1	102	412	70	4
2	2	39	322	36
3	1	2	70	166

Figure 5. Essay set 4

Actual marks	0	1	2	3	4
Predictions					
1	19	189	63	1	0
2	0	54	370	74	1
3	0	1	93	336	74
4	0	0	1	28	140

Figure 6. Essay set 5

Actual marks	0	1	2	3	4
Predictions					
1	30	96	21	1	0
2	0	39	241	58	0
3	0	3	71	505	86
4	0	0	0	94	195

Figure 7. Essay set 6

Actual marks	0- 2	3- 5	6- 8	9-11	12-14	15-17	18-20	21-23	24-26
Predictions									
6- 8	0	0	1	0	0	0	0	0	0
9-11	1	6	60	49	34	6	0	0	0
12-14	0	2	19	68	113	83	12	0	0
15-17	0	0	1	12	54	174	109	13	1
18-20	0	0	0	0	9	77	89	53	16
21-23	0	0	0	0	2	8	34	52	35
24-26	0	0	0	0	0	1	8	29	21
27-29	0	0	0	0	0	0	0	1	2

Figure 8. Essay set 7

Actual marks	6-11	12-17	18-23	24-29	30-35	36-41	42-47	48-53	54-59
Predictions									
18-23	1	0	2	0	0	0	0	0	0
24-29	0	1	2	16	12	1	0	0	0
30-35	0	0	1	15	102	29	2	0	0
36-41	0	0	0	0	75	193	51	2	0
42-47	0	0	0	0	4	25	29	13	0
48-53	0	0	0	0	0	0	1	0	1

Figure 9. Essay set 8

Scores on one word essays

Here we provide confusion matrices on validation set for each of 8 LSTM models, one for each essay set.

Score	1	2	3	4	5	6	7	8	9	10	11	12
# of essays	0	10	995	1665	**4241**	1270	1375	1351	1226	990	1420	457

Table 4. Distribution of scores received by one-word essays. Model for essay set 1.

Score	1	2	3	4	5	6
# of essays	**2847**	**4758**	1626	1045	828	**2896**

Table 5. Distribution of scores received by one-word essays. Model for essay set 2.

Score	0	1	2	3
# of essays	272	**2066**	510	**3152**

Table 6. Distribution of scores received by one-word essays. Model for essay set 3.

Score	0	1	2	3
# of essays	**1137**	974	172	**1717**

Table 7. Distribution of scores received by one-word essays. Model for essay set 4.

Score	0	1	2	3	4
# of essays	386	**1395**	254	265	**1700**

Table 8. Distribution of scores received by one-word essays. Model for essay set 5.

Score	0	1	2	3	4
# of essays	558	**1769**	237	259	**2177**

Table 9. Distribution of scores received by one-word essays. Model for essay set 6.

Score	0-3	4-6	7-9	10-12	13-15	16-17	19-21	22-24	25-27	28-30
# of essays	**2027**	970	662	**2043**	352	216	221	312	622	**2360**

Table 10. Distribution of scores received by one-word essays. Model for essay set 7.

Score	0-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50	51-55	56-60
# of essays	30	367	557	825	**1371**	978	756	548	621	**1105**	**2571**	**1271**

Table 11. Distribution of scores received by one-word essays. Model for essay set 8.

Adversarially perturbed essays

Original grade: **4** out of 6

Grade after modifications: **5** out of 6

In @DATE1's ~~world~~ **planet**, there are ~~many~~ **numerous** things ~~found~~ **detected** offensive. Everyone has their own opinion ~~on~~ **concerning** what is offensive and what is not. Many parents are becoming ~~upset~~ **outraged** because they think their children are viewing things that they should not. Other ~~people~~ **citizens** are ~~upset~~ **outraged** because they think the libraries are offending their culture or ~~way~~ **paths** of life. This is even taken ~~to~~ **of** the extreme ~~where~~ **thus** ~~people~~ **citizens** want censhorship ~~on~~ **concerning** libraries ~~to~~ **of** avoid this, which is ~~wrong~~ **misguided**. Some ~~people~~ **citizens** are becoming concerned ~~about~~ **toward** the materials ~~in~~ **at** libraries. They find these things ~~to~~ **of** be offensive. Everyone is entitled ~~to~~ **of** their own opinion, but there really is nothing anyone can do ~~if~~ **whether** ~~someone~~ **person** is offended.

The ~~world~~ **planet** is a public place and everywhere we go, something might be ~~found~~ **detected** offensive. The ~~library~~ **librarians** is a place for study. It is never attended ~~to~~ **of** offend ~~someone~~ **person**, or bring bad ~~to~~ **of** the ~~world~~ **planet**. It is ~~simply~~ **sheer** a place ~~to~~ **of** inform, and ~~if~~ **whether** ~~someone~~ **person** is offended by what they ~~see~~ **admire**, they should stay away from the ~~library~~ **librarians**. I have been ~~to~~ **of** the ~~library~~ **librarians** ~~many~~ **numerous** times, none of which have I ever seen anything offensive. Everything I have ever witnessed at the ~~library~~ **librarians** is for learning and research. There are certain sections ~~in~~ **at** the ~~library~~ **librarians**. If a parent does ~~no~~ **neither**t want their ~~child~~ **childhood** seeing something, they should keep their ~~child~~ **childhood** ~~in~~ **at** the children's section. I can ~~assure~~ **ensures** ~~you~~ **thou**, there is nothing offensive ~~in~~ **at** the children's section, or else the ~~library~~ **librarians** ~~would~~ **ought** not have it ~~in~~ **at** that section. The owners of these libraries know what is going ~~to~~ **of** ~~upset~~ **outraged** ~~people~~ **citizens** and what will not. If there was truly offensive materials ~~in~~ **at** the ~~library~~ **librarians**, those materials ~~would~~ **ought** be taken out. Also, ~~if~~ **whether** a person complains, and the materials are ~~removed~~ **eradicated**, it ~~could~~ **would** lessen ~~someone~~ **person** else's chance getting the materials they need. One person ~~could~~ **would** think the material is offensive, but ~~someone~~ **person** else might want ~~to~~ **of** learn more ~~about~~ **toward** it. If one is offended by a certain material, all they ~~simply~~ **sheer** must do, is not look at it. The ~~library~~ **librarians** can be compared ~~to~~ **of** a big computer. One can basically find anything there. Asking the ~~library~~ **librarians** ~~to~~ **of** censor their materials is like asking the internet ~~to~~ **of** censor theirs. It is a ~~way~~ **paths** of learning and researching and it ~~would~~ **ought** be almost impossible ~~to~~ **of** censor everything there. Everyone is going ~~to~~ **of** be offended some point ~~in~~ **at** their life. If the libraries ~~removed~~ **eradicated** everything that ~~could~~ **would** offend ~~someone~~ **person**, they ~~would~~ **ought** have ~~no~~ **neither** materials left. People need ~~to~~ **of** stop being so easily offended and ~~realize~~ **reaching** the ~~library~~ **librarians** is not ~~trying~~ **striving** ~~to~~ **of** harm anyone. There does not need ~~to~~ **of** be any censorship ~~in~~ **at** libraries. It is ~~simply~~ **sheer** ~~trying~~ **striving** ~~to~~ **of** teach ~~people~~ **citizens** ~~about~~ **toward** the ~~world~~ **planet** and let them enjoy books, music, movies, or whatever else one might go ~~to~~ **of** the ~~library~~ **librarians** ~~to~~ **of** find.

Original grade: **3** out of 6

Grade after modifications: **4** out of 6

Yes and no, some materials such as books, music, movies, magazines, etc, ~~should~~ **must** be ~~voted~~ **adopted** upon the citizens ~~to~~ **of** be removed from shelves. I do think that some materials in those catagories ~~should~~ **must** be removed ~~if~~ **whether** they are ~~offensive~~ **abusive** ~~to~~ **of** me as well as others, but it will take a long while ~~to~~ **of** get them removed from stores and other places ~~if~~ **whether** other people ~~like~~ **fond** them. I do not ~~like~~ **fond** how they make some music ~~to~~ **of** be very violent and ~~cause~~ **provoke** ~~minds~~ **souls** of most teenagers ~~to~~ **of** ~~turn~~ **transform** ~~bad~~ **wicked** and start selling ~~drugs~~ **drug** on the street of their hometown, but i can't do anything about that because that kind of music is admired by those teenagers as well as some adults too. If some people can buckle down and see that stuff ~~like~~ **fond** that will ~~mess~~ **chaos** up lives of teenagers and some adults who fall victim ~~to~~ **of** it, then there is a chance that it can be stopped. Stopping ~~things~~ **elements** ~~like~~ **fond** this will ~~save~~ **rescue** a community from disaster and ~~cause~~ **provoke** other good chances in life ~~for~~ **into** people in need ~~for~~ **into** those chances. Here's ~~another~~ **further** example, ~~like~~ **fond** this music artist named @CAPS1 @CAPS2. She has made some ~~great~~ **wonderful** songs ~~for~~ **into** the past year and a half now. People have ~~told~~ **say** me that she is part of a group called @CAPS3 and its a group ~~where~~ **hence** they try ~~to~~ **of** I think 're-birth' thereselves. My ~~friends~~ **boyfriends** wanted me ~~to~~ **of** stop listening ~~to~~ **of** her music. I ~~told~~ **say** my ~~friends~~ **boyfriends** that I do not ~~like~~ **fond** the fact that she joined this group, but that doesnt mean im gonna stop listen ~~to~~ **of** her music. Now ~~if~~ **whether** she makes a song that is ~~offensive~~ **abusive** ~~to~~ **of** me and as well as my friends, then that ~~where~~ **hence** I draw the line. What im saying is that ~~if~~ **whether** people don't have others ~~to~~ **of** back them up ~~if~~ **whether** something is highly ~~offensive~~ **abusive** ~~to~~ **of** them and oblivious ~~to~~ **of** others, it will be very hard trying ~~to~~ **of** prove yourself in the best ~~way~~ **manner** possible.

Original grade: **3** out of 6

Grade after modifications: **4** out of 6

Would ~~you~~ **thou** want your childern reaing about ~~things~~ **subjects** that only @CAPS1 know's what? When ~~you~~ **thou** go ~~to~~ **of** a library ~~you~~ **thou** aspect ~~to~~ **of** learn about @CAPS2, @CAPS3, @LOCATION1's @CAPS4,@CAPS5 etc. Libraries ~~are~~ **constitute** for learning new ~~things~~ **subjects** about the world that will ~~later~~ **subsequently** ~~help~~ **helps** ~~you~~ **thou** in @CAPS9. When ~~you~~ **thou** first walk into a library ~~you~~ **thou** except ~~to~~ **of** see people checking out @CAPS2 books , or books that catch your eye ~~just~~ **merely** by the title. If we find a book is offensive , or will not ~~help~~ **helps** better our childerns' future ~~then~~ **upon** stand up and fight for their own mental development. We must ~~also~~ **likewise** think of what the childern want ~~to~~ **of** read. They have the right ~~to~~ **of** read what ever they want, as long as it's ~~entertaining~~ **hilarious** ~~to~~ **of** them and they ~~are~~ **constitute** learning something new. Some books teaches them about the world they ~~are~~ **constitute** growing up in. There ~~are~~ **constitute** some books that I would not let my own ~~child~~ **childhood** read, but I know in my

heart that she is learning something that I @MONTH1 not be able ~~to~~ **of** teach her. Those types of books of ~~are~~ **constitute** what I call, '@CAPS7's'. Those books that can come off seeming offensive, when in the end they ~~are~~ **constitute** actually, what I ~~call~~ **appealed** '@CAPS8', helping ~~to~~ **of** ~~prepare~~ **elaborate** them for what is it come. Not every book will be full of rainbows, ~~pretty~~ **rather** colors, or pop-ups. They must know that they ~~are~~ **constitute** some people they have ~~to~~ **of** be mindfull of, and people who ~~are~~ **constitute** educating them on @CAPS9. They have ~~to~~ **of** learn the difference between what's right, and what's wrong. Remember the first book ~~you~~ **thou** ever read by yourself? I do. It was called 'Of @CAPS10 and @CAPS11'. I read that book when I was @NUM1. Till this day my mother says, 'I tried ~~to~~ **of** stop ~~you~~ **thou** from reading that book ~~so~~ **therefore** many times , it ~~had~~ **has** ~~dangerous~~ **dangers** wording that an @NUM1 year ~~should~~ **ought** no be able ~~to~~ **of** read at that young age'. What she did not know was that; that book ~~had~~ **has** taught me alot about the world back then. That knowledge I ~~had~~ **has** obtain ~~then~~ **upon** ~~had~~ **has** helped me ~~later~~ **subsequently** on my @CAPS9. Some books ~~are~~ **constitute** ment ~~to~~ **of** be read while some aren't. If ~~you~~ **thou** feel your ~~child~~ **childhood** ~~should~~ **ought** not read a certain book ~~then~~ **upon** read it for yourself, and ~~then~~ **upon** tell your ~~child~~ **childhood** the reason why they can not read the same book ~~you~~ **thou** ~~had~~ **has** ~~just~~ **merely** read.

Artificial essays by generative RNN

Essay written by student. Score: **5** out of 6. The highlighted text fragment is used as a seed for generative RNN.

How @CAPS4 you feel if your favorite book was taken off the shelves of your school or public library? I, along with many other students, @CAPS4 find this discouraging and distastrous, so I do not believe that censorship should affect books that are on the shelves. Otherwise, a demolished love of reading, crushed individuality, and separated population @MONTH1 be born. Like the beloved @PERSON2 @PERSON2 series by @PERSON1, many books and series are being taken out of libraries' collections due to people in society finding them offensive. In this case, the world of witchcraft in which this story blooms is against some religious beliefs; therefore, some individuals within a religion campaign to have these books banned. Fortunately, none of the libraries I visit, with their eclectic collections, had banned this series, or I @CAPS4 not have the strong thirst for literature as I do now. All books have the potential to pull a student into the wonderful world of reading, like @PERSON2 did for me, so taking away books that are most likely to spark an interest or start a firework of creativity @CAPS4 not only affect this generation, but the futures of all. If this censorhip was to be allowed, who is to say what all could be censored? Who @CAPS4 be the final judge as to what books @CAPS4 be banned? It @CAPS4 all come down to power and who was willing enough to take it. This struggle to be on top has the possibility of seperating people apart like political parties. Disagreements could turn into debates, and those could turn into fights. It can be concluded that people are stubborn for their beliefs, and to have someone choose what everyone is allowed to believe @CAPS4 be wrong. For instance, it @CAPS4 be like an @CAPS1 forcing a @CAPS2 to not believe in @CAPS3; a vegetarian commanding that meat can no longer be eaten; a woman taking away men's voting rights. Censorship @CAPS4 lead to the disrespect of other's opinions, and disrespect is never a beneficial thing. Each and every person has a different opinion on what is offensive or not, so to censor books @CAPS4 be to censor all individual mentality. Without each person's unique thoughts and beliefs, the world @CAPS4 become similiarly vapid and dull. Differences in beliefs is what adds variety to the population and what makes a person special; additionally, free thought is a right all people should have. If someone was to limit the mental, literary stimulants that are out in the world, the amount of creativity and individuality @CAPS4 decrease. To conclude, censorship @CAPS4 be a disrespect to individuality, personal beliefs, and the overall joy of reading a good book. Just because one might not believe in what a story says, it does not mean that the piece of literature should be forbidden. No one is being forced to read the books that grace the hundreds of shelves in a library, so if someone is offended, simply do not read it. So how @CAPS4 you feel if your favorite book was gone from all libraries? Disrespected? That is how I @CAPS4 feel

Essay written by RNN. $T = 1$. Score: **4** out of 6

how caps you feel if your favorite book was taken off the shelves of your school or public library ? i , along with many other students , caps find this discouraging and distastrous , so i do not believe that there should be being removed from the shelf because they find offensive and . caps that we have the children have been taking away people down for simply things . caps caps caps caps shelf ' and freedom of time ? a story is to enjoy a gender , is a create example , day , and walls , fiction being about they . magazines should be censorship in the today violates their as . adult for one look students , / on the pieces of not if the they make more like . caps caps , our is some just have some they open it were then they do not be censored that removed about it , not the caps1 why or month month with this magazine because there were another for a certain whole and book . reading the library , a what you of those public , location , from the novel . caps take a children in censored , everyone warning will be removed from the shelves , so they would not have to listen for . strong , there would n't

something that your big from life simply it by not walk . the library is how into the caps goes , because they would know it case to watching it reader , deal , or see the i offensive . when this comes the reading , i think that the caps caps caps month n't think to those power that enters those way to the movies , in their young or take your child n banned they like . if a major well , or unique how down works unconstitutional , true , and liberty through . many caps , people in caps1 watching an an keep music for caps1 of one world not young take something in the information 's everything will provided right . what does n't . we can still every your our school and can not these books are being type and . take might yelling what is once should not be sexual at the has : into series or find a certain book and very own nation . such as books should not be censored , because i find offensive , she should get not story if the caps 's caps right to be a had ways to novel for the public of the point . nothing for should to be censored . such as people and put , say , i , but you take in libraries should that the novel i question what they get to read an 's free time and they know all books about a taken of caps time ? , reading books i allowed my allowed to read about what they might

Essay written by RNN. $T = 10$. Score: **5** out of 6

how caps you feel if your favorite book was taken off the shelves of your school or public library ? i , along with many other students , caps find this discouraging and distastrous , so i do not believe handful supported traumatized lost sexually view unsatisfactory resulted derrogatory risky must signs harmed pre relationships minutes an concept becomes drink lessons rid further wither early band losing took horrid destination collide plain religous hav energy grain interests arisen feet immoral completely ears enlightenment friends quoted knowing bannig thier theirlives almost priority ordinance dealing forbidden violate first prompting desperately among date present starts industries take friends mentality secondly him critical recomneded awaiting means collectively around perfectly frustrating unlike dive sentiment progression play forbiden reality similiarly but complicated fine aspects selected blind want easily decisions persistance task expressed vary worlds food will mental discussing offenseive determined burning defeated surrs brand bare negative oftentimes survuve artist explicate princes signature horrid ironically unconstitutional readers malignant houses closed knowlege gears parts grab benefit introduce conflict ways shops remving tragic 'lost reported careful created thus aflame given belong happens reason tactic 'kid historian arises all grow necessary truely comes these efficient commented picture wild fond carefully populaion unalienable guidance grow inapropriate mother somebody no shown threaten 'unsuitable 'censorship course favorite falling on process psychological thrilled dish resistance act magazine bawdy story listen that main would rent during lessening publicly easily thrown dissapointed playing spill vile behave lose safeguards attempts language 'right boundaries outlook sorted anywhere kind tale possibility liked denied deal significant stick blurred 'safety intailed determine everyone fairly perceived plagued supposedly difficult remembering n't both wrestling stolen thinking glance to major nothin objects fear intolerable lovers religous argue ordinance accentuated fed modest devastated interesting difference daughter shows entitled helped libary offending source singer racist eyes rose symbol experiences imagaintion referring undertone negative nurtured wrong viewed she children meanings banned common oneself inspired despite when habit interpret nearly visited try revealing putting owned watch possibly removing hurts dimensions grave wack boarding harmonies a complain list founded tend storage 'mistake students chosen stolen shadowed document ventures innappropriate plan view divide utilizes gore references included cruel thousands are enlightenment such public trilogy kids although wrong denied sources incinerated view furthermore among risks in unfortunately particular says politcal lead finish believes centuries everybody sold receive sixteen gives white ask inventors occurance easily decisions affecting factor clearly innicent thier death standards thing put punish dubbed ideals forcing enter incident suited wish rappers back adventures healthy begining rebelious value slowly promotes progression beauty offer worldwide boy curiosity off bundled reads accptable encourage walks on videogame religions disgrace appropriate least alow freiends beings choir 'find novel draws fits governments civilization amendments effort intrest frequently blowing tell safeguards due cruel proudly cut lost an locked complaining bible morally happening refers dispute vanished inject regarding desires matter divine opportunity but added gangs troubling certain opposite understands i whomever dispose producers discard empower about restricting disturbed 'you aflame ties offensive innappropriate opinions enveloped million discretion