Automated Essay Scoring: Natural Language Processing

Parth Tamane
ptamane3@gatech.edu

Gaurav Pande
gpande@gatech.edu

Naila Fatima
nfatima3@gatech.edu

Abstract

Automated essay scoring (AES) is a long term goal of the natural language processing world. This field has gained traction as NLP made progress in sentiment analysis. Many researchers claim that due to the subjective nature of an essay there is a bias associated with its grading, and this bias can be due to many factors including the person who is grading, gender of the author, location of the author, etc. Due to this reason, we would like to create models which can grade essays objectively.

1 Motivation

Essay writing is a standard language practice and having a human manually grade these essays can be a difficult task as it both tedious and subjective. Organizations such as ETS (Attali and Burstein, 2014) already incorporate automated methods to perform the essay scoring task. However, we can use machine learning models (such as SVM and regression) as well as deep learning models (such as LSTM, RNN and Highway networks) (Salvatore et al., 2003) to improve the grading mechanism and understand the various biases attached to manual grading like gender bias. (Tony et al., 2019)

1.1 Literature Survey

Automated grading can be achieved either using style based analysis, content based analysis or a mixture of both these techniques. (Salvatore et al., 2003) With content-based analysis, we are more interested in what an essay says. (Kakkonen et al., 2005)(Kakkonen et al., 2006) We represent each essay with dense vectors (obtained by decomposing word-document context matrices) and using cosine similarity to assign grades. Another approach to learn from content is to use n-gram embeddings with CNNs (Dong and Zhang, 2016) and LSTMs (Taghipour and Ng, 2016). Style-based features (such as those used by the PEG grader: essay length, word length, etc.) can be used understand the syntactic structure of the text (Page, 1994). This approach can be useful when more open ended prompts need to be graded. Finally, (Larkey, 1998) uses both types of features to train a linear regression model. First, this method uses content based features with a Bayesian classifier to predict the probability that the essay is good. It also uses averaged k-nearest neighbours scores as another feature. Finally, 11 style based features like number of sentences, unique words, etc are also used. Both feature types are given to a linear regression model. In our project we want to explore with these different models defining new stylistic features and augmenting old features used. We will be using the different methods based on their appropriateness for the different types of essays. Our initial hypothesis is that content based techniques will work better on less open ended prompts while style based techniques would work better on more open ended prompts.

2 Goal

There are 2 main goals of the project. The first goal is to try and train models that will give a higher correlation score with human graded scores. The second goal is to analyze the essays and check for trends in scores with that of the author characteristics like gender, age, etc. Together these two goals will help us get a holistic understanding of how essays are graded by humans and get new insights which can help future efforts in automatic essay grading.

3 Plan

3.1 Dataset

We will be using the ”The Hewlett Foundation: Automated Essay Scoring” dataset for our project.

(Hewlett, 2012: accessed March 12, 2020) This data set consists of 8 essay prompts with over 13,000 transcribed essays and ratings. The prompts can be divided in to 2 distinct categories: 1) Persuasive/Narrative/Expository (opinion on a topic) and 2) Source dependent responses (comprehension of a given text). The dataset also contains grades assigned by different graders on varying scales. For profiling task, we will be using a Kaggle dataset (J. Schler and Pennebaker, 2004: accessed March 13, 2020) which includes blog posts along with the gender and age of the bloggers.

3.2 Models

One of our goal is to use deep learning methods to address the AES problem by training deep learning models on approximately 13000 essays with their respective scores. There has already been a lot of research done in this field but mostly machine learning techniques have been used. The paper by Tanghipour and Ng (Taghipour and Ng, 2016) was one of the earliest paper involving the idea of convolution to automated scoring. Drawing inspiration from this, we would like to experiment with RNNs and CNNs. The intuition behind this is that RNNs can learn complex structures and embedding in the data, and therefore can be useful in deriving meaningful features that can be used for scoring.

We will also be experimenting with linear regression based approach on content and style driven features. While we don't have the resources, datasets or the time to engineer and update best features on a per prompt basis continuously as done by ETS for their e-rater, (Salvatore et al., 2003) (Attali and Burstein, 2014) we feel that more handpicked features can improve the essay rating performance. We plan to build on the 11 features chosen in (Larkey, 1998) and include features that measure word beauty, (Shihui Song, 2013: accessed March 12, 2020) correct word count, correct sentence count, etc.

We intend to find out if characteristics associated with a certain demographic (gender, age, etc) are preferred over those associated with a different demographic. Demographic information tends to embed itself in the identities of an individual. It is possible to extract features which capture information related to the contextuality and formality of the written text as well as the grammatical structure of the sentences used. Features such as F1-measure, POS-sequence patterns, n-gram models and word embedding are able to capture information related to gender (Rao et al., 2010) (Mukherjee and Liu, 2010) and age (Rao et al., 2010). We can use models such as logistic regression and support vector machines (SVM) to find patterns in the data. Since most of the research pertaining to author profiling has been conducted on social media posts or blogs, we will be training our classifiers on these datasets. We assume that since essay-writing is generally a time-constrained exercise (most tests require a student to complete an essay under an hour), essays will still display features which embed the demographic information of the writer.

3.3 Timeline

For the midterm report, we plan to work on feature selection. Also, we will work on visualizing the data to understand how the features correlate to essay scores. We also plan on training a model which can predict the gender of the author. For the final report, we plan to train models which can score an essay using machine learning and deep learning approaches. We intend to us Google Colab as training and evaluating platform for models, and as backup if the models fails to give proper result we will fall back to basic models (like Linear Regression, SVM). We will also create a model which can predict the age of the author and use the predicted demographic information to outline any biases which may be present in the data.

3.4 Commitment

We are planning to distribute the tasks in the following manner: All three of us will work on feature extraction and visualization. Parth will be focusing on machine learning techniques like linear regression based model on style and content based features. Naila will be focusing on author profiling to derive biases and other features from the essay. Gaurav will be focusing on deep learning based techniques for training and evaluating essay scores.

3.5 Evaluation

We will be using Cohen's kappa score (Cohen, 2020: accessed March 12, 2020) method, a widely used evaluation technique to calculate the correlation between score assigned by 2 raters(one human rater and other machine rater). We use correlation because essay rating is a very subjective task. Hence, comparing the accuracy based on absolute scores won't be a very fair evaluation method.

References

Yigal Attali and Jill Burstein. 2014. Automated essay scoring with e-rater® v.2.0.

Cohen. 2020: accessed March 12, 2020. *Cohen's Kappa*.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring – an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1072–1077, Austin, Texas. Association for Computational Linguistics.

Hewlett. 2012: accessed March 12, 2020. *The Hewlett Foundation: Automated Essay Scoring*.

S. Argamon J. Schler, M. Koppel and J. Pennebaker. 2004: accessed March 13, 2020. *Blog Posts Labeled with Age and Gender*.

Tuomo Kakkonen, Niko Myller, and Erkki Sutinen. 2006. Applying latent dirichlet allocation to automatic essay grading. *Lecture Notes in Computer Science*, 4139:110–120.

Tuomo Kakkonen, Niko Myller, Jari Timonen, and Erkki Sutinen. 2005. Automatic essay grading with probabilistic latent semantic analysis. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pages 29–36, Ann Arbor, Michigan. Association for Computational Linguistics.

Leah S. Larkey. 1998. Automatic essay grading using text categorization techniques. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 90–95, New York, NY, USA. Association for Computing Machinery.

Arjun Mukherjee and Bing Liu. 2010. Improving gender classification of blog authors. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

Ellis Batten Page. 1994. Computer grading of student prose, using modern concepts and software. In *The Journal of Experimental Education 62, no. 2*, pages 127–42.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents - SMUC 10*.

Valenti Salvatore, Neri Francesca, and Alessandro Cucchiarelli. 2003. An overview of current research on automated essay grading. *Journal of Information Technology Education*, 2.

Jason Zhao Shihui Song. 2013: accessed March 12, 2020. *Automated Essay Scoring Using Machine Learning*.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Tony, Gaut, Andrew, Tang, Huang, Yuxin, Mai, Zhao, Mirza, Elizabeth, and et al. 2019. Mitigating gender bias in natural language processing: Literature review.