HIGH SCHOOL EDITION
Journal of Student Research

Volume 13 Issue 1 (2024)

Automated College Application Essay Grading

Rinah Zhang[1], Stepan Malkov[#] and Peter Mbua[#]

[1]Holmdel High School, USA
[#]Advisor

ABSTRACT

In recent years, computer programs used to score essays have been explored extensively, with many different approaches being developed. Most of these approaches use Natural Language Processing (NLP) techniques (Ade-Ibijola et al., 2012), a field of machine learning often used to analyze and understand text. These approaches fall under the name of Automated Essay Scoring (AES), which typically assesses essay quality with a single score (Ke and Ng, 2019). This paper proposes a natural language processing (NLP) model which predicts the quality of a college application essay, which is proximally measured through a college's acceptance rate. Key essay factors include the number of grammar mistakes, sophistication of writing, repetition, and the text of the essay. Multiple different models were tested. A Random Forest Classifier relying solely on grammar, sophistication of writing, and repetition metrics achieved the best performance, yielding an accuracy of 89.7%. The second-best model was a combination of an LSTM and a logistic regression model. Other models significantly underperformed, yielding accuracies in the range of 40%-60%. Ultimately, our model may help a number of students going through the college application process to understand where their essay may stand compared to other students.

Introduction

In the past few years, there has been an explosion of machine learning applications in the field of natural language processing (NLP) (Otter et al., 2021). Key examples include part-of-speech tagging, language translation (Nadkarni et al., 2011), or summarizing a piece of text (Liddy, 2001). Another application of NLP is automated essay scoring (AES), which grades responses by considering certain grammatical and linguistic features of the text (Ramesh and Sanampudi, 2021). The first version of such an AES, Project Essay Grader (PEG), was developed in 1966. It evaluated the key writing characteristics of the text: content, organization, style, mechanics, etc (Page and Paulus, 1968). More recently, these systems began using regression-based and NLP techniques due to better performance compared to previous versions (Ramesh and Sanampudi, 2021). Motivated by the growing field of literature incorporating NLP techniques into essay grading, this work aims to develop a version of an AES tailored to college application essays. In the 2022-2023 US college application cycle, 1,244,476 applicants applied through the Common Application (a common college application form used in undergraduate admissions), and this number is predicted to increase in upcoming years (Nietzel, 2023). One of the biggest pieces of one's application are the essays, which include the personal statement and additional supplemental essays (Munson, 2021). A plethora of resources/programs are out there to help students edit their college essays, but access to such resources can prove to be very expensive (Jaschik, 2019). Considering the apparent disparities in affordability and how many students go through the application process every year, this paper proposes an NLP model that judges essay quality. This model could later be turned into a service to help serve as a baseline for essay quality and editing for students of all socioeconomic classes. The remainder of the paper is structured as follows: Section 2 presents the methodology for data collection and preprocessing, while Section 3 discusses model architecture and training. Section 4 presents the results of the different models tested, as well as an analysis of the limitations of the model. Section 5 concludes the paper.

Data Collection and Preprocessing

To build this model, data was collected from a variety of sources. The data was in the form of successful college application essays admitted into certain colleges. (See Table 3 for the list of accessed data sources). Other data points collected were the college acceptances of a particular essay, as well as that college's acceptance rate. All the acceptance rates were determined according to U.S. & News World Report's listings and were recorded into a spreadsheet. These colleges were then split into different tiers (see Table 1) based on their acceptance rate. These tiers can be described as Extremely Competitive, Very Competitive, Competitive, Somewhat Competitive, Slightly Competitive, and Not Competitive.

Table 1. Acceptance rate tiers and corresponding acceptance rate ranges

Tier	Acceptance Rate Range (%)
1	0-5
2	6-10
3	11-20
4	21-40
5	41-60
6	61-100

To measure essay quality, we considered three key parameters: grammar, sophistication, and repetition. (See Table 2 below for the metrics and how they were measured). These measurements and all the machine learning and regression architecture were implemented with the help of Python's Keras and Tensorflow packages.

Table 2. Key features of an essay used as metrics for the NLP model.

Metric	Form of Measurement
Grammar	Number of Mistakes
Sophistication	Gunning Fog Index
Repetition	Frequency of Adjectives

Grammar

Writing that is full of grammar and spelling mistakes may often hurt acceptance chances (Cairns, 2022). To measure essay grammar quality, we propose to use the number of grammatical mistakes made throughout the entire essay. To measure this, the package python_language_tool was used: a Python wrapper for LanguageTool, which finds the number of grammar/spelling mistakes in a piece of text.

Sophistication

For sophistication, which is a much more subjective metric, we rely on the Gunning Fog Index. This is an index that measures the readability of a piece of writing. It estimates how many years of formal education are needed to understand the piece of text on the first reading, through objective features. (Scott, 2023) For instance, an index of 6 represents a sixth-grade level, a 12 represents a high school senior level, and a 17 is the level of a college graduate (Notorc, 2006). The formula for the Gunning Fog Index is (Scott, 2023):

$$0.4(ASL + PHW)$$

where ASL is the average sentence length and PHW is the percent hard words (hard words: words that are not proper nouns which contain 3 or more syllables, a combination of one-syllable/hyphenated words, or three-syllable verbs that with an -es or -ed ending.) While the Gunning Fog Index has its flaws, it is a simple enough quantitative measure of sophistication that can be easily interpreted by a machine learning model (DiMAscio, n.d). gunning_fog is a function under the Python Readability package that can provide the score that a piece of text receives on the index.

Repetition

Repetition of words can be seen as undesirable, especially if it is excessive. (Fanning, 2012) This is why repetition was included as another metric for the model to consider. Each word was counted by how many times it was repeated in the text (not including common words like "the" or "and"). Using nltk's part of speech tagging package, Repetition analysis was limited to just adjectives. The reason only adjectives are considered is because certain words like nouns, especially proper nouns, are often necessary to the text, and often have no substitutes. On the other hand, adjectives have multiple synonyms and thus act as a better proxy for the level of sophistication for a given piece of text. Correspondingly, the function adj_frequency, will return how many times the most frequent adjective used appeared throughout the text.

Text Cleaning

Finally, the cleaning of the text to prepare for the model was done. Essays were cleaned by hand, making sure no errors were transferred over during copy and pasting, and then put into an Excel spreadsheet. The next step was to split up the text. Tokenization is a frequent practice used in NLP to split text into smaller units (or tokens) that can then be assigned a meaning, whether these units are characters, words, phrases, etc (Burchfiel, 2022). In the case of this work, the text was chosen to be split into words. This process helped prepare the data for the machine learning model. Each essay was split into sentences using spaCy. Then, all the words were lower cased, and stop-words were removed (stop-words are common words like "the" or "and"). These words were identified using nltk's (Natural Language Toolkit) stopwords function. Since on average, a sentence is between 15-20 words in length, (Gina, n.d), Each sentence was padded to 15 words using padding tokens to create a uniform size of data while leaving most of the data in. Then, every 5 sentences were combined (75 tokens) until there were no more sentences left. If the last few sentences did not add up to 5, they were ignored to avoid too many padding tokens.

All these metrics - grammar, sophistication, repetition, and text vectors - were added into a data frame alongside the school accepted, the acceptance rate, and the school's acceptance rate tier. Since each essay was split into multiple distinct inputs, each measure was evaluated for the entire text and then repeated for each input.

Architecture and Training

Data analysis can be performed in several ways, such as using statistical techniques or machine learning techniques. Machine learning models can prove more powerful in providing accurate predictions for more complex relationships in comparison to statistical models (Turintech, n.d). This is why a machine learning model was chosen for this paper's purposes. All machine learning models take vectors as an input, so when it comes to text, it needs to be converted into numerical form. To do this, an embedding layer using Keras's Embedding package was created. This embedding layer converts each word into an unique integer. Similar words are encoded by proximal integers (TensorFlow, 2023). In this embedding layer, a dimension of 10,000 was used, which resulted in 7.638 million parameters.

One common network used in NLP tasks and other types of sequential data is a Long Short-Term Memory (LSTM) network, which is a type of Recurrent Neural Network (RNN). Like its name suggests, LSTMs add and remove past data based on what it deems important. It does this by having 3 different gates - an input gate, a forget gate, and an output gate. These gates filter data to be added or removed from the layer's memory (Chugh, 2023). Given its proficiency in processing text data, an LSTM layer was added to the model. This layer consisted of 64 units. Model architectures with multiple LSTM layers were also tested, but the performance of the model did not improve.

Lastly, the model incorporated a dense layer with an output of one out of the 6 tiers of acceptances (See Figure 1).To train the model, hyperparameters were set as follows: epochs were set to 100, batch size was set to 32, and the validation split was 80%-20%. The output of this model (a number between 1-6) was then used as an input, along with the text's number of grammar mistakes, sophistication level, and repetition metrics for a logistic regression model, with the output remaining one of the 6 acceptance tiers.

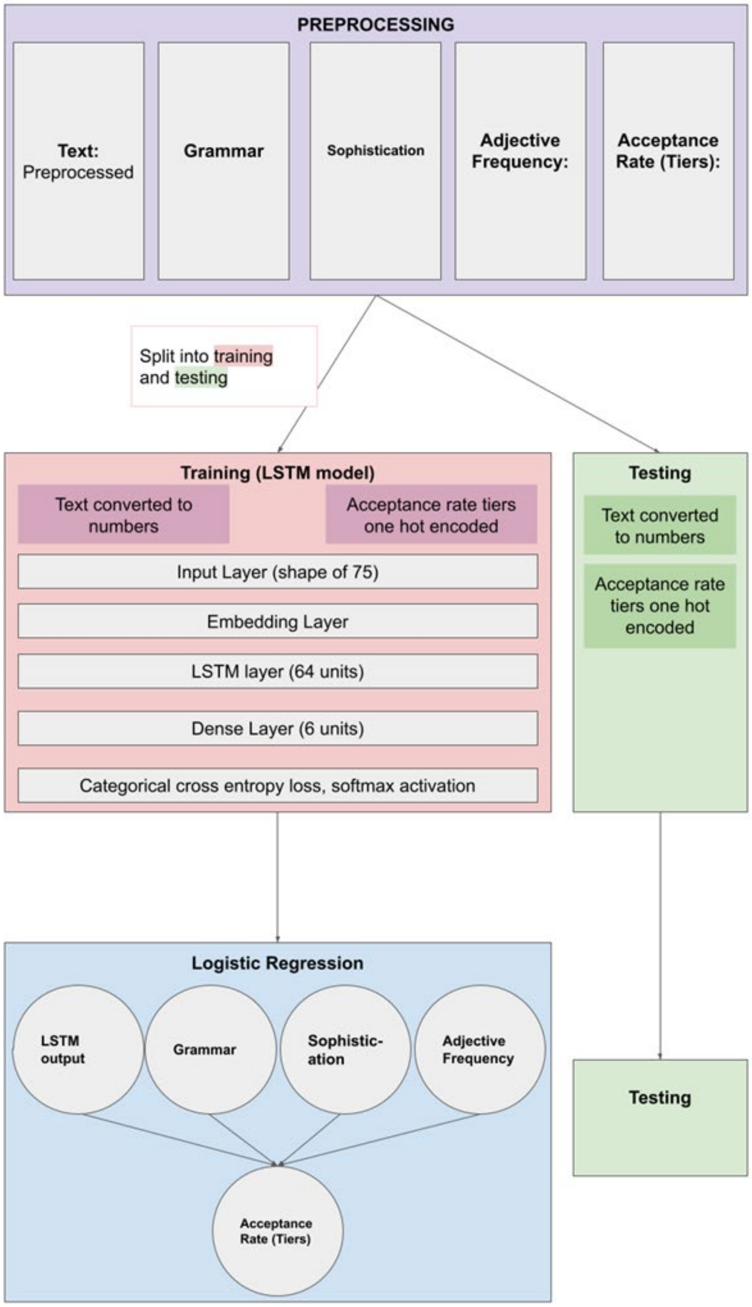

Figure 1. LSTM + Logistic Regression Model Architecture. Key model features and steps taken. The model follows from text preprocessing to a combined LSTM-logistic regression model that predicts an acceptance tier for a given essay. Other steps include splitting the data into training and testing, and one hot encoding of acceptance tiers.

Variations of the original LSTM-regression model which were additionally tested included a pure LSTM model, a logistic regression model (without the text), an LSTM and random forest classifier model, and lastly a random forest classifier without the text of the essays.

Discussion and Limitations

A total of 5 models with varying architectures were tested (See Table 3 for the different models and their accuracies). A pure LSTM model using text as input produced an accuracy of 52.38%, while a combined LSTM-logistic regression module using grammar, sophistication, and repetition as additional inputs resulted in an accuracy was 53.06%. The same inputs used in a Random Forest Classifier gave back an accuracy of 43.23%. Given the minimal increase from the LSTM model to the logistic regression, there's reason to assume that the LSTM model did the bulk of the work. That's why a logistic regression model without the LSTM output was implemented, which had an accuracy of 41.29%. The last model tested was a Random Forest Classifier without the LSTM output which resulted in a total accuracy of 89.68%.

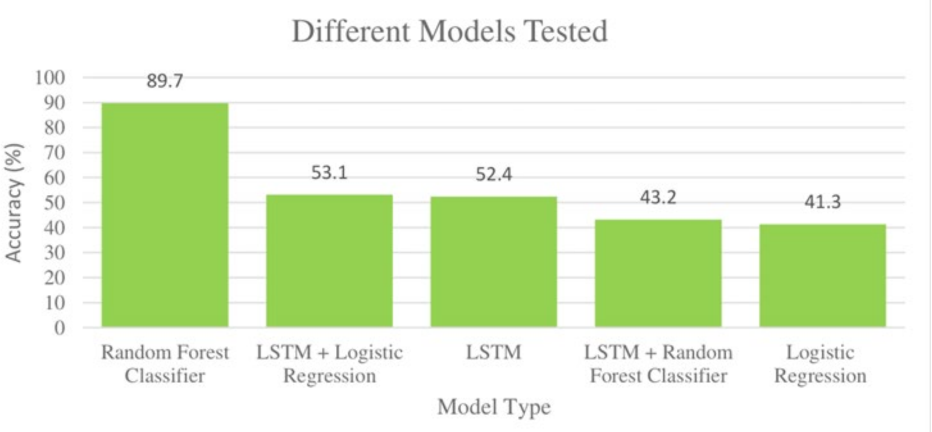

Figure 2. Different Model Types. Bar graph depicting the different accuracies for the 5 models tested.

Table 3. Model Types and Accuracy. Type of models tested, and their resulting accuracies.

Model Type	Accuracy (%)
LSTM	52.4
LSTM + Logistic Regression	53.1
Logistic Regression	41.3
LSTM + Random Forest Classifier	43.2
Random Forest Classifier	89.7

There are a few plausible reasons why the Random Forest Classifier model's accuracy was so high compared to the other models. The metrics chosen, grammar, sophistication, and repetition might have been informative on their own. Another reason could be that the linguistic variability of the essays was too difficult for the LSTM model to generalize effectively. This variability could be a result from the different prompts and topics that the essays responded to. This then calls into the question why the logistic regression performed poorly when compared to a random forest. Random Forests assign varying importance to features, which then allows to root out unimportant features (Kirasich et al., 2018). Additionally, logistic regression assumes relationships between variables and the target are linear. Particularly, Random Forests capture non-linear interactions better than logistic regression (SaturnCloud, 2023).

Additionally, there are many factors taken into consideration for a college application, such as extracurriculars, recommendation letters, awards, etc. (Munson, 2021) aside from just the essays. This project only considers the essay component of a student's college application, but it is clear from previous machine learning studies (see Lee et al., 2023) that even when a single application factor, such as a student's demographics, is dropped, the accuracy of the model also drops (by around 7%).

Another limitation to this study is the ever-changing nature of the college admissions process. For example, in the past two decades, applications have increased by more than 150%. (Selingo, 2022) This means that essays from a few years ago may not have been held to as high a standard as they are now. Thus, models trained on historical data may become less accurate in the future.

Lastly, an issue with the data is the level of quality gathered. Most of the websites that these essays were gathered from were created to serve as references and examples for people applying to certain colleges. This means, they are probably posting essays of high quality, even if it's a lower tier school. Due to the small sample size of lower quality essays in the data collected, it is unclear if the model can accurately grade these essays.

Conclusion

From the different models tested, the Random Forest Classifier worked the best with an accuracy of 89.68%, but if one wants the model to consider the actual substance of the text, a combination of an LSTM model and logistic regression yields the second-best performance, with an accuracy of 52.38%. It is important to also consider that there is a plethora of other factors considered in college admissions. The low degree of accuracy for the LSTM model potentially suggests that, although important, the essay is not the only thing determining the acceptance rates of these colleges. With this, future research in this field could focus on the inclusion of additional application factors, like GPA, extracurriculars, etc. into the model to achieve higher model accuracies and yield more accurate predictions.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

Burchfiel, A. (2022, May 16). What is NLP (Natural Language Processing) Tokenization? - tokenex. TokenEx. https://www.tokenex.com/blog/ab-what-is-nlp-natural-language-processing-tokenization/

Cairns, H. (2022, December 22). 4 Things Admissions Officers Don't Like In Your College Essays. College Raptor. https://www.collegeraptor.com/getting-in/articles/college-applications/4-things-in-college-app-essays-that-admissions-officers-dont-like/

Chugh, A. (2023, May 31). Deep Learning | Introduction to Long Short Term Memory. GeeksforGeeks. https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory/

DiMAscio, C. (n.d.). py-readability-metrics · PyPI. PyPI. https://pypi.org/project/py-readability-metrics/

Fanning, P. (2012, April 7). 24. Good & Bad Repetition. guinlist. https://guinlist.wordpress.com/2012/04/07/24-good-bad-repetition/

Gina. (n.d.). How Many Words Should Be in a Sentence? LanguageTool. Retrieved September 27, 2023, from https://languagetool.org/insights/post/sentence-length/

Jaschik, S. (2019, March 17). A look at the many legal ways wealthy applicants have an edge in admissions. Inside Higher Ed. https://www.insidehighered.com/admissions/article/2019/03/18/look-many-legal-ways-wealthy-applicants-have-edge-admissions

Ke, Z., & Ng, V. (2019, July). Automated Essay Scoring: A Survey of the State of the Art. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 6300-6308. 10.24963/ijcai.2019/879

Kirasich, K., Smith, T., & Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. SMU Data Science Review, 1(3). https://scholar.smu.edu/datasciencereview/vol1/iss3/9

Lee, J., Thymes, B., Zhou, J., Joachims, T., & Zizilcec, R. F. (2023, June 30). Augmenting Holistic Review in University Admission using Natural Language Processing for Essays and Recommendation Letters. https://doi.org/10.48550/arXiv.2306.17575

Liddy, E. D. (2001). Natural Language Processing. In Encyclopedia of Library and Information Science (2nd ed.). N.Y. Marcel Decker, Inc.

Munson, N. (2021, December 13). 9 basic elements of a complete college application. Marymount University. https://marymount.edu/blog/9-basic-elements-of-a-complete-college-application/

Nadkarni, P. M., Machado, L. O., & Chapman, W. W. (2011, Sep-Oct). Natural language processing: an introduction. J Am Med Inform Assoc, 18(5), 544-51. 0.1136/amiajnl-2011-000464

Nietzel, M. T. (2023, March 30). College Applications Are Up Dramatically In 2023. Forbes. https://www.forbes.com/sites/michaeltnietzel/2023/03/30/college-applications-are-up-dramatically-in-2023/

Notorc. (2006, November 2). Clear Writing: How to Achieve and Measure Readability. Postscripts. http://notorc.blogspot.com/2006/09/devils-in-details-measuring.html

OluAde-Ibijola, A., Wakama, I., & Amadi, J. (2012). An Expert System for Automated Essay Scoring (AES) in Computing using Shallow NLP Techniques for Inferencing. International Journal of Computer Applications, 51, 37-45. 10.5120/8080-1480

Otter, D. W., Medina, J. R., & Kalita, J. K. (2019, December 21). A Survey of the Usages of Deep Learning of Natural Language Processing. IEEE Transactions on Neural Networks and Learning Systems, 32(2), 22. https://doi.org/10.48550/arXiv.1807.10854

Page, E. B., & Paulus, D. H. (1968, April). The Analysis of Essays by Computer. Final Report. ERIC. https://eric.ed.gov/?id=ED028633

Ramesh, D., & Sanampudi, S. K. (2021, September 23). An automated essay scoring systems a systematic literature review. Artif Intell Rev, 55(3), 2495-2527. 10.1007/s10462-021-10068-2

SaturnCloud. (2023, July 10). Why Is Logistic Regression Not Working But Decision Tree Is? Saturn Cloud. https://saturncloud.io/blog/why-is-logistic-regression-not-working-but-decision-tree-is/

Scott, B. (2023, August 1). The Gunning's Fog Index (or FOG) Readability Formula – ReadabilityFormulas.com. Readability Formulas. Retrieved September 27, 2023, from https://readabilityformulas.com/the-gunnings-fog-index-or-fog-readability-formula/

Selingo, J., & Mull, A. (2022, March 23). The College-Admissions Process Is Completely Broken. The Atlantic. https://www.theatlantic.com/ideas/archive/2022/03/change-college-acceptance-application-process/627581/

Sorensen, T. (2022, April 18). What Parents of College Applicants Need to Know. USNews.com. https://www.usnews.com/education/blogs/college-admissions-playbook/articles/changes-that-parents-of-college-applicants-need-to-know

TensorFlow. (2023, May 27). Word embeddings | Text. TensorFlow. https://www.tensorflow.org/text/guide/word_embeddings

Turin Tech. (2021, October 4). Machine Learning vs Statistical Modelling: which one is right for your business problem? – TurinTech AI. TurinTech AI. Retrieved October 1, 2023, from https://www.turintech.ai/2021/10/04/machine-learning-vs-statistical-modelling-which-one-is-right-for-your-business-problem/