Automated Essay Scoring

Joaquin Farina

Abstract

Automatic Essay Scoring (AES) systems are a powerful tool that improves the efficiency and fairness of the grading process and support writing skills' acquisition for students. It can also become extremely useful as a support tool for teachers.

Nevertheless, the technological aspect of developing an accurate and robust AES remains largely a challenge. In this project, I create a machine-learning based AES for the Independent Writing Assignment of the TOEFL test.

Based on language development knowledge and research-informed feature engineering, I create a model that achieves an accuracy of up to XX% for the XXX class. Strengths and weaknesses, as well as potential improvements are also discussed.

Automated Essay Scoring and Machine Learning

In the context of automatic essay scoring, machine learning algorithms can be trained to recognize patterns and features in written essays that are indicative of good writing quality.

Using machine learning techniques, automatic essay scoring systems can learn to identify features such as grammar, spelling, and punctuation errors, as well as more complex features such as organization, structure, and the use of evidence and examples. The system can then use this knowledge to evaluate and grade essays based on these features.

Overall, the use of machine learning in automatic essay scoring allows for more accurate and efficient grading of written work, as well as the ability to provide immediate feedback to students and identify areas for improvement in their writing skills.

Evaluation Methods

This project is heavily based on machine learning algorithms. Models are fitted using training data (label and covariates), and subsequently are evaluated using test data, by comparing predicted values to the actual values (ground truth) set aside for evaluation. Five evaluating standards are used in this assignment:

I. **Accuracy** = (True Positives + True negatives) / (Positives + Negatives)

II. **Recall** = True Positives / (True Positives + False Negatives)

III. **Precision** = True Positives / (True Positives + False Positives)

IV. **Cohen's kappa** = (p_o - p_e)/(1-pe), p_o = empirical probability of agreement, p_e = expected probability of agreement.

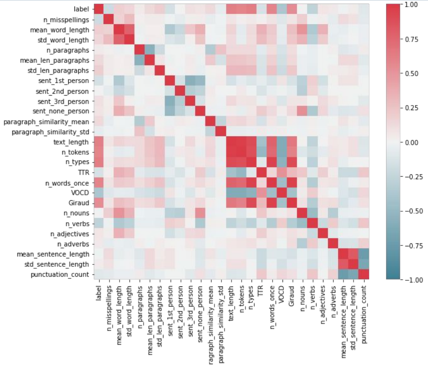
V.

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\frac{\mathrm{precision}\cdot\mathrm{recall}}{\mathrm{precision}+\mathrm{recall}} = \frac{2\mathrm{tp}}{2\mathrm{tp}+\mathrm{fp}+\mathrm{fn}}.$$

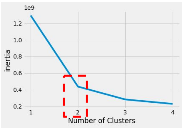
Features

1. Misspelled words
2. Average and std word length
3. Number of paragraphs, and mean-std length
4. Point of view (# first / second / third person)
5. Similarity between paragraphs (mean and std)
6. Text length
7. Number of tokens
8. Number of unique tokens
9. TTR
10. Number of words occurring once
11. VOCD # TTR = [D*/N][1+2N/D)½ - 1] >>> solve for D
12. Giraud
13. Percentage of nouns, verbs, adj, adv
14. Sentence length (mean, std)
15. Word length (mean, std)
16. Count punctuation

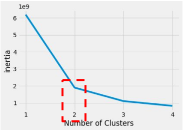
Correlations



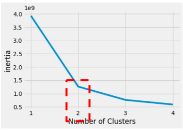
Clustering



Low

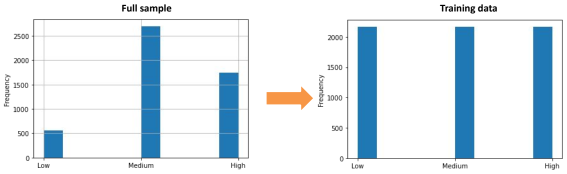


Medium



High

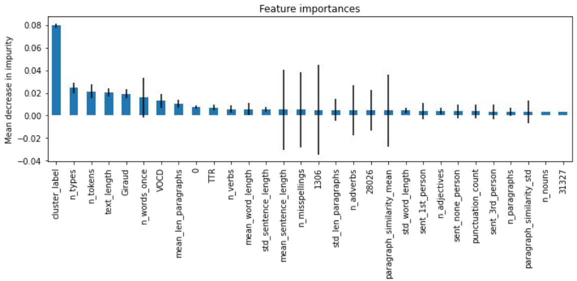
Oversampling



Results

	SVM	LDA	RF	KNN	NB	NB*
Accuracy	0.64	0.55	0.83	0.59	0.91	0.944
Precision	0.65	0.59	0.84	0.6	0.91	0.945
Recall	0.64	0.55	0.83	0.59	0.91	0.944
F1-score	0.64	0.56	0.83	0.59	0.91	0.943
Kappa	0.57	0.45	0.7	0.5	0.88	0.93

Feature importance



Feature: text similarity with respect to high scoring essays (mean of similarities).

Contact

Joaquin Farina
Teachers College, Columbia University
jef2177@tc.columbia.edu