Automated Essay Scoring in the Presence of Biased Ratings

Evelin AmorimMarcia CançadoAdriano VelosoComputer Science DepartmentLinguistics DepartmentComputer Science DepartmentUFMGUFMGUFMGBrazilBrazilBrazilevelin@dcc.ufmg.brmcancado@ufmg.bradrianov@dcc.ufmg.br

Abstract

Studies in Social Sciences have revealed that when people evaluate someone else, their evaluations often reflect their biases. As a result, rater bias may introduce highly subjective factors that make their evaluations inaccurate. This may affect automated essay scoring models in many ways, as these models are typically designed to model (potentially biased) essay raters. While there is sizeable literature on rater effects in general settings, it remains unknown how rater bias affects automated essay scoring. To this end, we present a new annotated corpus containing essays and their respective scores. Different from existing corpora, our corpus also contains comments provided by the raters in order to ground their scores. We present features to quantify rater bias based on their comments, and we found that rater bias plays an important role in automated essay scoring. We investigated the extent to which rater bias affects models based on hand-crafted features. Finally, we propose to rectify the training set by removing essays associated with potentially biased scores while learning the scoring model.

1 Introduction

Automated Essay Scoring (AES) aims at developing models that can grade essays automatically or with reduced involvement of human raters (Page, 1967). AES systems may rely not only on grammars, but also on more complex features such as semantics, discourse and pragmatics (Davis and Veloso, 2016; Song et al., 2014; Farra et al., 2015; Somasundaran et al., 2014). Thus, a prominent approach to AES is to learn scoring models from previously graded samples, by modeling the scoring process of human raters. When given the same set of essays to evaluate and enough graded samples, AES systems tend to achieve high agreement levels with trained human raters (Taghipour and

Ng, 2016).

While research in AES has focused on designing scoring models that maximize the agreement with human raters(Chen and He, 2013; Alikaniotis et al., 2016), there is a lack of discussion on how biased are human ratings. Despite making judgments on a common dimension, raters may be influenced by their attitudes, their cultural background, and their political and economic views (Guerra et al., 2011). Since AES models are designed to learn by analyzing human-graded essays, AES models could inherit rating biases present in the scores from human raters, and this may result in systematic errors. Thus, our objective in this paper is to examine the extent to which rater bias affects the effectiveness of stateof-the-art AES models. A deeper understanding of such factors may help mitigating the effects of rater bias, enabling AES models to achieve greater objectivity.

In order to study the effects of rater bias in essay scoring, we created an annotated corpus containing essays written by high school students as part of a standardized Brazilian national exam. Our corpus contains a number of essays, written in Portuguese, along with their respective scores. Further, raters must also provide a comment for each essay in order to ground their scores. As in (Recasens et al., 2013) we built subjectivity and sentiment lexicons that serve as features to represent the comments, that is, rater comments are represented according to the subjectivity distribution as given by specific subjectivity cues in our lexicons. We present empirical evidence suggesting that the subjectivity distribution within rater comment is a proxy for the score that is given to the essay. More specifically, very low (or very high) scores are associated with essays for which rater comments showed a very particular subjectivity distribution. We also investigated the relationship between subjectivity distribution and the misalignment between human raters and AES models. Interestingly, the subjectivity distribution becomes very characteristic as the misalignment increases.

Our main contributions are three-fold:

- We built subjectivity lexicons for the Portuguese language. These lexicons include words and phrases associated with different subjectivity dimensions – sentiments, factive verbs, entailments, intensifiers and hedges. We identify biased language within rater comments by calculating the word mover's distance (Kusner et al., 2015) between comments and the lexicons. This approach benefits from large unsupervised corpora, that can be used to learn effective word embeddings (Mikolov et al., 2013). By identifying biased language, we observed that biases can work to inflate essay scores or to deflate them.
- We employ a set of linguistic features in order to learn different AES models, and we evaluate the effects of biased ratings in the efficacy of these models. In summary, biased ratings affect AES models in different ways, but in general the misalignment between human rater and the AES model is more acute when the rater shows biased language in their comments.
- We propose simple ways of preventing and reducing the negative effects of biased ratings while learning AES models. Results in a controlled experimental setting revealed that detecting and removing biased ratings from the training set lead to significant improvements in automated essay scoring.

In the remainder of this paper, Section 2 discusses related work on automated essay scoring. Section 3 describes the features used for learning AES models, as well as the features used for identifying biased language in rater comments. Further, our debiasing approach is also discussed in Section 3. Section 4 describes the data, the setup and the results of our empirical evaluation. Finally, Section 5 provides our conclusions.

2 Related Work

Research in cognitive science, psychology and other social studies offer a great amount of work on (conscious and unconscious) biases and their effects on a variety of human activities (Kaheman and Tversky, 1972; Tversky and Kaheman, 1974). Biases can create situations that lead us to make decisions that project our experiences and values onto others (Baron, 2007; Ariely, 2008). While there is sizeable literature on rater effects in general settings (Myford and Wolfe, 2003), it remains unknown how biased ratings affect automated essay scoring models. Rather, works on automated essay scoring are mainly focused on designing AES models by maximizing the agreement with human raters, despite the assertiveness of the ratings.

Typically, AES systems are built on the basis of predefined linguistic features that are then given to a machine learning algorithm (Amorim and Veloso, 2017). Works that fall into this approach include (Srihari et al., 2008, 2007; Cummins et al., 2016; McNamaraa et al., 2015). Further, authors in (Dong and Zhang, 2016) presented an empirical analysis of features typically used for learning AES models. Authors in (Crossley et al., 2015) studied a broader category of features that can also be used to build AES models. There are also more recent approaches for learning AES models that do not assume a set of predefined features. These approaches are based on deep architectures, and include (Alikaniotis et al., 2016; Taghipour and Ng, 2016; Riordan et al., 2017; Dong et al., 2017). Finally, there also models based on domain adaptation (Phandi et al., 2015) and unsupervised learning (Chen et al., 2010).

Few works have investigated the subjective nature of essay scoring. An interesting exception is (Allen et al., 2015), in which the authors investigated the misalignment between students' and teachers' ratings of essay. Results revealed that students who were less accurate in their self-assessments produced essays that were more causal, contained less meaningful words, and had less argument overlap between sentences.

The work in this paper builds upon prior work on building subjectivity lexicons (Klebanov et al., 2012) and subjectivity detection (Recasens et al., 2013), but in our case applied to score agreement. In this respect, our work is more comparable to (Klebanov and Beigman, 2009; Beigman and Klebanov, 2009), where authors discussed and investigated the problem of learning in the presence of biased annotators. Other works that are also close to ours include (Farra et al., 2015; Somasundaran et al., 2016; Song et al., 2014), in which the authors studied the problem of scoring persuasive and argumentative essays.

3 Method

Our aim in this work is to learn AES models that are less prone to the effects of biased ratings, that is, models that are able to perform highly objective and impartial judgements. Thus, we start this section by proposing features that are useful for building AES models. Then, we propose another set of features that are useful for identifying biased ratings based on subjectivity cues. Finally, we propose an approach to remove biased ratings from the training set, thus learning more objective AES models.

3.1 Features for Essay Scoring

As most existing AES systems, our models are built on the basis of predefined features (e.g. number of words, average word length, and number of spelling errors) that are given to a machine learning algorithm. The features used to build our AES models are discussed and evaluated in (Amorim and Veloso, 2017). They may fall into two broad categories:

- **Domain features:** These are simple linguistic features, including the number of first-person pronouns, demonstrative pronouns and verbs. Features also include the number of pronouns and verbs normalized by the number of tokens in the corresponding sentence.
- **General features:** Most of the general features are based on (Attali and Burstein, 2006). However, due to lack of tools for processing the Portuguese language, we implemented the following features, which are sub-divided as follows:
 - **Grammar and style:** Features include the number of grammar errors and misspellings. These numbers are also normalized by the number of tokens in the corresponding sentence. In order to evaluate style, we designed features based on the style rules suggested in (Martins, 2000). Features include the number of style errors and the number of style of errors per sentence.

Organization and development: Features

include the number of discourse markers from the Portuguese grammar, and the number of discourse markers per sentence. Discourse markers are linguistic units that establish connections between sentences to build coherent and knit discourse.

Lexical complexity: Features include the Portuguese version for the Flesh score (Martins et al., 1996), the average word length (i.e., the number of syllables), the number of tokens in an essay, and the number of different words in an essay.

Prompt-specific vocabulary usage:

Features include different distances between prompt and essay (i.e., cosine distance). In this case, both the prompt and the essay are treated as frequency vectors of words.

3.2 Features for Identifying Biased Ratings

We assume a scenario in which essay raters must ground the provided scores with specific comments. We also assume that we can identify biased ratings by detecting comments with biased language. In order to detect biased language, we developed subjectivity lexicons for the Portuguese language. Specifically, a linguist built a list of Portuguese lexicons based on the analysis of expressions that seem to express some subjectivity of the human evaluator. Our subjectivity lexicons are categorized into the following groups:

- Argumentation: This lexicon includes markers of argumentative discourse. Argumentative markers include lexical expressions and connectives, such as: "even" (*até*), "by the way" (*aliás*), "as a consequence" (*como consequência*), "or else" (*ou então*), "as if" (*como se*), "rather than" (*em vez de*), "somehow" (*de certa forma*), "despite" (*apesar de*), among others.
- **Presupposition:** This lexicon includes markers that suggest the rater assumes something is true. Some examples of such markers include: "nowadays" (*hoje em dia*), "to keep on doing" (*continuar a*), and factive verbs.
- **Modalization:** This lexicon indicates that the writer exhibits a stance towards its own state-

ment. Some examples of such markers are adverbs, auxiliary verbs, modality clauses, and some type of verbs.

- **Sentiment:** This lexicon also includes markers that indicate a state of mind or a sentiment of the rater while evaluating the essay. Some examples of such markers include: "with regret" (*infelizmente*), "with pleasure" (*felizmente*), and "it is preferable" (*preferencialmente*).
- **Valuation:** This lexicon assigns a value to facts. Usually, adjectives are employed as valuation, but as adjectives are context dependent we use only in this class the markers related to intensification, such as: "absolutely" (*absolutamente*), "highly" (*altamente*), and "approximately" (*aproximadamente*).

3.3 Debiasing the Training Set

Bias is generally defined as a deviation from a norm. If the norm is unknown to us, then bias is hard to identify. Thus, our approach for debiasing the training set starts by finding the norm (in terms of the subjectivity within rater comments) for each score value. Intuitively, the amount of subjectivity within a comment should be similar to the amount of subjectivity within another comment, given that the scores associated with the corresponding essays are close to each other. So, we should not expect to find essays having discrepant scores, but for which the corresponding comments show a similar amount of subjectivity. Our debiasing approach is divided into three steps:

1. Rater comments are represented according to the amount of subjectivity cues. In order to represent a comment, we calculate the distance between it and each of the five subjectivity lexicons. More specifically, we learn word embeddings (Mikolov et al., 2013) for the Portuguese language, and then we employed the Word Mover's Distance function (Kusner et al., 2015) between a comment and the five subjectivity lexicons. As a result, each comment is finally represented by a five-dimensional subjectivity vector, where each dimension corresponds to the amount of a specific type of subjectivity. This results in a subjectivity space, where comments are placed according to their amount of subjectivity.

- 2. We group subjectivity vectors according to the score misalignment associated with the corresponding essay. Then, we calculate centroids for each group in order to find the prototypical subjectivity vector for each group (or misalignment level).
- 3. The distance to the prototypical subjectivity vector is used as a measure of deviation from the norm. Specifically, we sort essays according to the distance between the subjectivity vector and the corresponding centroid. Then, we define a number of essays to be removed from the training set. The relative number of essays to be removed from the training set is controlled by hyper-parameter α .

4 Experiments

In this section, we present the data we used to learn and evaluate different AES models. Then, we discuss our evaluation procedure and report the results obtained with our debiasing approach. In particular, our experiments aim to answer the following research questions:

- **RQ1:** How scores are distributed across the essays? How aligned with human raters are different AES models?
- **RQ2:** Does subjectivity in rater comments vary depending on the given score?
- **RQ3:** Does subjectivity in rater comments vary depending on the misalignment between the AES model and the human rater?
- **RQ4:** Can we mitigate the effects of biased ratings?

4.1 Corpus

Our corpus is composed of essays (n = 1, 840) that were written by high-school students as part of a standardized Brazilian national exam. Each essay is evaluated according to the following five objective aspects:

- Formal language: Mastering of the formal Portuguese language.
- **Relevance to the prompt:** Understanding of essay prompt and application of concepts from different knowledge fields, to develop the theme in an argumentative dissertation format.



Figure 1: Distribution of the scores given by human raters. Also, distribution of the scores given by different AES models.

- **Organization of information:** Selecting, connecting, organizing, and interpreting information.
- **Argumentation:** Demonstration of knowledge of linguistic mechanisms required to construct arguments.
- **Solution proposal:** Formulation of a proposal to the problem presented.

The final score is given as the sum of the scores associated with each aspect. Raters are supposed to perform impartial and objective evaluations, and they must enter specific comments in order to ground their scores. Also, each essay was assessed by one rater.

Bias-free ratings: We also separate a number of essays (n = 50) which received similar scores by three expert raters who were directly instructed to perform impartial, objective, and unbiased evaluations. These raters are PhD-level in Linguistics with unlimited time to provide their ratings, and they do not participate on the creation of the training set. We assume the ratings given to these essays were not contaminated by biased judgements, and we will use these essays for evaluating the efficacy of AES models learned after the training set is debiased.

4.2 Setup

We implemented the different AES models using scikit-learn (Pedregosa et al., 2011). Specifically, we learn AES models using Support Vector Regression (SVR), Random Forests (RF), Logistic Regression (LR), Gradient Boosting (GB), and Multi-Layer Perceptron (MLP). All models are based on the same set of features, previously



Figure 2: Distribution of misalignment for the different AES models.

described in Section 3.1, and all models are trained in regression mode. The measure used to evaluate the effectiveness of the different models is the quadratic weighted kappa (κ) which measures the inter-agreement between human raters and AES models (Cohen, 1960). We conducted five-fold cross validation, where the dataset is arranged into five folds with approximately the same number of examples. At each run, four folds are used as training set, and the remaining fold is used as test set. We also kept a separate validation set. The training set is used to learn the models, the validation set is used to tune hyper-parameters and the test set is used to estimate κ numbers for the different the models. Unless otherwise stated, the results reported are the average of the five runs, and are used to assess the overall effectiveness of each model. To ensure the relevance of the results, we assess the statistical significance of our measurements by comparing each pair of models using a Welch's ttest with p-value ≤ 0.01 .

4.3 Results and Discussion

Next we report results obtained from the execution of the experiments, and discuss these results in the light of our research questions.

Score distribution: The first experiment is concerned with RQ1. Figure 1 shows how scores are distributed over the essays in our corpus. Although the distribution differs for each AES model, scores are centered around 4, and few essays received extreme scores. The LR model seems to have a preference for lower scores. The scores provided by the GB and MLP models are better distributed.

Figure 2 shows how aligned with human raters are the different AES models. For most of the essays, AES models are well aligned with human



Figure 3: Subjectivity distribution for human raters.

raters, showing misalignments that vary from -2 to +2. For some essays, the LR model tends to give scores that are much smaller than the score given by the human rater. The GB and MLP models perform very similary, but the MLP model shows a slightly better alignment.

Subjectivity vectors and biased ratings: The second experiment is concerned with RQ2. Figure 3 shows the average subjectivity vector grouped according to the score given to the corresponding essay (i.e., the centroid or prototypical vector of a score). More specifically, we first grouped subjectivity vectors according to the score associated with the corresponding essay, and then we calculated the average subjectivity vector for each group. As shown in Figure 3, the argumentation dimension increases with the score, while modalization tends to decrease. Presupposition, valuation and sentiment dimensions show a very similar trend with varying score values.

Figure 4 shows t-SNE representations (van ter Maaten and Hinton, 2008) for the average subjectivity vectors (centroids for each group of score). Three larger clusters emerged: subjectivity vectors associated with score 0, subjectivity vectors associated with scores between 1 and 6, and subjectivity vectors associated with scores between 6 and 10.

Subjectivity vectors and misalignment: The third experiment is concerned with RQ3. Figure 5 shows the average subjectivity vector considering different levels of misalignment. More specifically, we grouped essays according to the misalignment between the score provided by the AES model and the human rater. Then, we calculated the average subjectivity vector for each group. As we can see, subjectivity affects AES



Figure 4: t-SNE representation for subjectivity vectors. Numbers correspond to the scores assigned to corresponding essays.

models in different ways. In general, however, subjectivity vectors within groups of essays associated with extreme misalignments are very different from subjectivity vectors associated with mild misalignments.

Figure 6 shows t-SNE representations for subjectivity vectors grouped by misalignment levels. Each cluster contains $\approx 80\%$ of the vectors associated with one of the misalignment levels inside the cluster. That is, 20% of the essays will be removed from the training set (i.e., $\alpha = 0.2$).

Debiasing the training set: The last experiment is concerned with RQ4. As described in Section 3.3, our debiasing approach works by removing from the training set a number of essays (controlled by α) that are more likely to be associated with biased ratings. Table 1 shows κ numbers for different α values. Clearly, the inter-agreement decreases as we remove essays with potentially biased ratings from the training set. This happens because the test set remains with essays that are potentially associated with biased ratings. In this case, removing biased ratings from the training set is always detrimental to the efficacy of AES models.

In order to properly evaluate our debiasing approach, we employ the 50 separate essays with bias-free ratings as our test set. In this case, biased ratings are removed from the training set, and the test set is composed by unbiased ratings. Table 2 shows κ numbers for different α values. As expected, the inter-agreement increases significantly with α , until a point in which keeping removing essays from the training set becomes detrimental. This happens either because we start to remove unbiased ratings, or the training set becomes too small. In all cases, the MLP model showed to be



Figure 5: Subjectivity distribution. (Top to bottom) SVR, RF, LR, GB, and MLP.



Figure 6: t-SNE representation for subjectivity vectors grouped by misalignment levels. The corresponding regions comprise essays associated with specific misalignment levels. (Top) GB model. (Bottom) MLP model.

	κ						
α	SVR	RF	LR	GB	MLP		
-	.404	.410	.408	.432	.446		
0.1	.390	.339	.364	.378	.393		
0.2	.365	.331	.344	.370	.393		
0.3	.345	.326	.338	.365	.386		
0.4	.340	.324	.333	.361	.384		
0.5	.307	.317	.328	.358	.382		

Table 1: κ numbers for different models with varying α values. There are potentially biased ratings in the test set.

	к						
α	SVR	RF	LR	GB	MLP		
-	.451	.472	.466	.491	.521		
0.1	.467	.491	.481	.505	.544		
0.2	.481	.511	.490	.521	.562		
0.3	.488	.526	.497	.542	.571		
0.4	.491	.523	.499	.547	.569		
0.5	.481	.518	.494	.545	.560		

Table 2: κ numbers for different models with varying α values. Ratings in the the test set are likely to be unbiased.

statistically superior than the other models.

5 Conclusions

In this paper, we investigated the problem of automated essay scoring in the presence of biased ratings. Most of the existing work on automated essay scoring is devoted to maximize the agreement with the human rater. This is fairly dangerous, since human ratings may be biased. Overall, discussion about the quality of the ratings in automated essay scoring is lacking, and this was a central interest in this paper. Specifically, we create a subjectivity space from which potentially biased scores/ratings can be identified. We showed that removing biased scores from the training set results in improved AES models. Finally, the essay data as well as the subjectivity lexicons that we will release as part of this research could prove useful in other bias related tasks.

Acknowledgments

This work was partially funded by projects InWeb (grant MCT/CNPq 573871/2008-6) and MASWeb (grant FAPEMIG/PRONEX APQ-01400-14), and by the authors individual grants from CNPq and FAPEMIG. AV thanks the support received from Kunumi.

References

- Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. pages 715–725.
- Laura Allen, Scott Crossley, and Danielle McNamara. 2015. Predicting misalignment between teachers' and students' essay scores using natural language processing tools. In *Proceedings of the 17th International Conference on Artificial Intelligence in Education.* pages 529–532.
- Evelin Amorim and Adriano Veloso. 2017. A multiaspect analysis of automatic essay scoring for brazilian portuguese. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Student Research Workshop)*. pages 94–102.
- Dan Ariely. 2008. Predictably Irrational: The Hidden Forces That Shape Our Decisions. HarperCollins.
- Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater (R) v. 2. *The Journal of Technology, Learning and Assessment* 4(3).
- Jonathan Baron. 2007. *Thinking and Deciding*, volume 4. Cambridge University Press.

- Eyal Beigman and Beata Beigman Klebanov. 2009. Learning with annotation noise. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*. pages 280–287.
- Hongbo Chen and Ben He. 2013. Automated essay scoring by maximizing human-machine agreement. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pages 1741–1752.
- Yen-Yu Chen, Chien-Liang Liu, Chia-Hoang Lee, and Tao-Hsing Chang. 2010. An unsupervised automated essay scoring system. *IEEE Intelligent Systems* 25(5):61–67.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1):37–46.
- Scott Crossley, Laura Allen, Erica Snow, and Danielle McNamara. 2015. Pssst... textual features... there is more to automatic essay scoring than just you! In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*. pages 203– 207.
- Ronan Cummins, Meng Zhang, and Ted Briscoe. 2016. Constrained multi-task learning for automated essay scoring. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. pages 789–799.
- Alexandre Davis and Adriano Veloso. 2016. Subjectrelated message filtering in social media through context-enriched language models. *Trans. Computational Collective Intelligence* 21:97–138.
- 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.
- Fei Dong, Yue Zhang, and Jie Yang. 2017. Attentionbased recurrent convolutional neural network for automatic essay scoring. In Proceedings of the 21st Conference on Computational Natural Language Learning. pages 153–162.
- Noura Farra, Swapna Somasundaran, and Jill Burstein. 2015. Scoring persuasive essays using opinions and their targets. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*.
- Pedro Henrique Calais Guerra, Adriano Veloso, Wagner Meira Jr., and Virgílio A. F. Almeida. 2011. From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pages 150– 158.
- Daniel Kaheman and Amos Tversky. 1972. Subjective probability: A judgment of representativeness. *Cognitive Psychology* 3(3):430–454.

- Beata Klebanov, Jill Burstein, Nitin Madnani, Adam Faulkner, and Joel Tetreault. 2012. Building subjectivity lexicon(s) from scratch for essay data. In *Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing*. pages 591–602.
- Beata Beigman Klebanov and Eyal Beigman. 2009. From annotator agreement to noise models. *Computational Linguistics* 35(4):495–503.
- Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*. pages 957– 966.
- E. Martins. 2000. *Manual de redação e estilo*. O Estado de São Paulo. https://books.google. com.br/books?id=CAkLnwEACAAJ.
- Teresa BF Martins, Claudete M Ghiraldelo, Maria das Graças Volpe Nunes, and Osvaldo Novais de Oliveira Junior. 1996. *Readability formulas applied to textbooks in brazilian portuguese*. Icmsc-Usp.
- Danielle McNamaraa, Scott Crossley, Rod Roscoe, Laura Allen, and Jianmin Dai. 2015. A hierarchical classification approach to automated essay scoring. *Assessing Writing* 23:35–59.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*. pages 3111–3119.
- C. Myford and E. Wolfe. 2003. Detecting and measuring rater effects using many-facet rasch measurement. *Journal Appl Meas*. 4(4):386–422.
- Ellis Page. 1967. Grading essays by computer: progress report. In *Proceedings of the Invitational Conference on Testing Problems*. pages 87–100.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in python. Journal of Machine Learning Research 12:2825–2830.
- Peter Phandi, Kian Ming Adam Chai, and Hwee Tou Ng. 2015. 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.

- Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. pages 1650–1659.
- Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. 2017. Investigating neural architectures for short answer scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. pages 159–168.
- Swapna Somasundaran, Jill Burstein, and Martin Chodorow. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of the 25th International Conference on Computational Linguistics*. pages 950–961.
- Swapna Somasundaran, Brian Riordan, Binod Gyawali, and Su-Youn Yoon. 2016. Evaluating argumentative and narrative essays using graphs. In *Proceedings of the 26th International Conference on Computational Linguistics*. pages 1568–1578.
- Yi Song, Michael Heilman, Beata Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the 1st Workshop* on Argument Mining. pages 69–78.
- Sargur Srihari, Jim Collins, Rohini Srihari, Harish Srinivasan, Shravya Shetty, and Janina Brutt-Griffler. 2008. Automatic scoring of short handwritten essays in reading comprehension tests. *Artif. Intell.* 172(2-3):300–324.
- Sargur Srihari, Rohini Srihari, Pavithra Babu, and Harish Srinivasan. 2007. On the automatic scoring of handwritten essays. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. pages 2880–2884.
- Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pages 1882–1891.
- Amos Tversky and Daniel Kaheman. 1974. Judgement under uncertainty: Heuristics and biases. *Science* 185:1124–1131.
- Laurens van ter Maaten and Geoffrey Hinton. 2008. Visualizing high-dimensional data using t-sne. Journal of Machine Learning Research 9:2579– 2605.