EXPLAINABLE MODELS FOR AUTOMATED ESSAY SCORING IN THE PRESENCE OF BIASED SCORING

EVELIN CARVALHO FREIRE DE AMORIM

EXPLAINABLE MODELS FOR AUTOMATED ESSAY SCORING IN THE PRESENCE OF BIASED SCORING

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação.

Orientador: Adriano Alonso Veloso
Coorientador: Márcia Cançado Lima

Belo Horizonte

16 de dezembro de 2019

EVELIN CARVALHO FREIRE DE AMORIM

EXPLAINABLE MODELS FOR AUTOMATED ESSAY SCORING IN THE PRESENCE OF BIASED SCORING

Thesis presented to the Graduate Program in Ciência da Computação of the Universidade Federal de Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Ciência da Computação.

Advisor: Adriano Alonso Veloso
Co-Advisor: Márcia Cançado Lima

Belo Horizonte
December 16, 2019

Amorim, Evelin Carvalho Freire de

A524e Explainable Models for Automated Essay Scoring in the Presence of Biased Scoring / Evelin Carvalho Freire de Amorim. — Belo Horizonte, 2019
xvi, 124 f. : il. ; 29cm

Tese (doutorado) — Universidade Federal de Minas Gerais — Departamento de Ciência da Computação.
Orientador: Adriano Alonso Veloso
Coorientador: Márcia Cançado Lima

1. Computação — Teses. 2. Aprendizado de Máquina — Teses. 3. Processamento de Linguagem Natural (Computação) — Teses. 4. Avaliação Automática de Ensaios. I. Orientador. II.Coorientador. III. Título.

CDU 519.6*82.10(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INCIPIT VITA NOVA
7 DE SETEMBRO DE 1927

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

EXPLAINABLE MODELS FOR AUTOMATED ESSAY SCORING IN THE PRESENCE OF BIASED SCORING

EVELIN CARVALHO FREIRE DE AMORIM

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

Adrian Alo Ve
PROF. ADRIANO ALONSO VELOSO - Orientador
Departamento de Ciência da Computação - UFMG

Márcia Cançado
PROFA. MÁRCIA MARIA CANÇADO LIMA - Coorientadora
Departamento de Linguística - UFMG

Fabrício Benevenuto
PROF. FABRÍCIO BENEVENUTO DE SOUZA
Departamento de Ciência da Computação - UFMG

PROF. PEDRO OLMO STANCIOLI VAZ DE MELO
Departamento de Ciência da Computação - UFMG

PROFA. CILENE APARECIDA NUNES RODRIGUES NEVINS
Departamento de Letras - PUC-Rio

Helena de M. Caseli
PROFA. HELENA DE MEDEIROS CASELI
Departamento de Computação - UFSCAR

Belo Horizonte, 16 de Dezembro de 2019.

Abstract

Written essays are the common way to select candidates for universities; therefore, students need to write essays as many as possible. Because of this, several methods for Automatic Essay Scoring (AES) for the English language have been proposed. Such methods should explain the score assigned to essays, and then the student can use the feedback to improve his or her writing skills. Therefore, many of the existing AES proposals employ handcrafted features instead of continuous vector representation. By using handcrafted features, it is easier for the system to give feedback to a student. Handcrafted features are also helpful to scrutinize the score assigned by an AES system and even the scores of human evaluators. This kind of investigation is useful to identify whether the features related to a writing skill are being considered during the assessment, which is essential if we desire fairer evaluations.

We present in this work an AES methodology to score essays according to five aspects or skills using handcrafted features and classical machine learning algorithms. In addition to that, we perform experiments to analyze which features influence which aspects in two different datasets evaluated by two distinct human evaluators. The performance of each aspect is explained by the feature analysis. Also, we explore the efficacy of AES models in the presence of biased data. Finally, we analyzed the evaluator's comments about essays by using a Portuguese lexicon list of biased words, which was assembled by Cançado et al. [2019].

Several experiments demonstrate the explainability of our models, and our proposed approach enhances the efficacy of AES models. The results regarding explainability are clear and assert that some features are particularly important for some aspects, while for other aspects, they are unimportant. We also show that the bias affects the efficacy of the classifiers, and when biased ratings are removed from the dataset, the accuracy of the model improves.

Palavras-chave: Machine Learning; Natural Language Processing; Automatic Essay Scoring.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Writing is a skill that evolves under the right guidance. Aware of such assumption, students around the world write several essays every week. Usually, specialized teachers are assigned to evaluate the produced essays, and students use teachers' feedback to improve their writing skills. However, reviewing essays is a time-consuming task and requires specific knowledge of evaluators.

Therefore, since the 1960s (Page [1994]), many attempts have been made to automate such a laborious task. Nowadays, frameworks that correct essays automatically are called Automatic Essay Scoring (AES) systems. Most of AES systems are for the English language and consider only one aspect in the evaluation process. As an aspect, we understand as a particular criterion, sometimes a subjective one, that evaluates an essay. One of the primary goals of this thesis is to develop models that can evaluate essays in a multi-aspect way. To achieve this goal, we will employ as the case study two datasets that are assessed according to the five aspects used by the National High School Exam (ENEM) of the Brazilian Government.

As explainability is a critical quality for an AES system, we investigate the influence of our features in the five aspects of our dataset. As it will be possible to observe, each aspect is affected in different ways by the handcrafted features. To investigate such influence, we employed ablation tests and a SHapley Additive exPlanations strategy(Lundberg and Lee [2017]).

In addition to that, we have the hypothesis that scoring is influenced by evaluator bias. This influence usually happens when a student declares an opinion that diverges from the opinion of the human evaluator. Therefore, we investigate the efficacy of AES strategies when there is biased scoring in the dataset, and we propose an unsupervised strategy to audit data under such circumstances.

1.1 Motivation

Written text is a proper way to evaluate people. Through an essay, it is possible to assess a person by his or her formal language knowledge, vocabulary knowledge, argumentative capacity, and many other aspects. However, evaluating essays is a task that demands a high human effort. When the essay is an official test, this is a critical issue, since more than a human evaluator is assigned to evaluate an essay to assure impartiality of scoring. The need for human resources and money is, then, doubled. In addition to that, every year, there are nearly seven millions enrollments for ENEM[1]. The cost to evaluate seven millions of essays every year is significant for the Brazilian government since two different human evaluators assess each dissertation. If there is a 20% percent difference between human evaluators grades, then a third evaluator is summoned to assess the dissertation.

Besides the charge of the national exam for the Brazilian government, due to the high workload, human evaluators can assess the same essay differently [Mendes, 2013] or even be affected by tiredness [Leckie and Baird, 2011], among other factors [Bridgeman, 2013]. Such heterogeneity in evaluations can result in different scores for similar essays, which leads to harmful consequences for some candidates. An automated scoring can aid in dealing with this issue by acting as a monitor of the evaluator's performance [Lottridge et al., 2013]. For instance, if the discrepancy between the scores assigned by the human and machine increases over time, then it is possible that the human is getting tired.

Another possible use for automatic scoring is the democratization of the writing practice for people of any social class. While a school with expensive teachers is a resource out of range for many students, they can easily download an application that automatic scores an essay or access a website that scores essays from their public school. This motivation is especially important when we consider that Brazil is a highly unequal country.

Even with the need for democratization of quality education and the existence of a big exam like ENEM in Brazil, few solutions were developed for the Portuguese language [Amorim and Veloso, 2017; Fonseca et al., 2018]. On the other hand, there are many proposals for AES frameworks in English [Larkey, 1998; Attali and Burstein, 2006; Chen and He, 2013; Alikaniotis et al., 2016; Tay et al., 2018]. Even though there are few AES in other languages [Zesch et al., 2015; Kakkonen and Sutinen, 2004], we aim to verify if such solutions are suitable for Portuguese.

[1] `http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/enem-2018-tem-6-7-milhoes-de-inscritos/21206`

Besides the language, another challenge that we have to face is the multi-aspect AES. None multi-aspect solution was developed for Portuguese to the best of our knowledge. Napoles and Callison-Burch [2015] introduced the Freshman Writing Corpus (FWC) of essays that are scored according to five aspects, which are: focus, evidence, organization, style, and format. In our research, we investigate the same set of aspects used in the ENEM evaluation. The FWC is similar to ENEM regarding the evaluation rubrics and is different concerning the formal aspect and the subjective differences described by the ENEM guidelines, such as to consider human values and social-cultural diversity in the solution proposal aspect. Then, we have to deal with an extremely subjectivity in the scores assigned to essays, which makes our task harder than other proposed AES tasks.

The explanation about the score assigned to an essay is also a crucial issue in this task. First, the auditing of a score is essential for the transparency of the scoring process. Second, feedback is necessary for a writer to improve. Hence, we analyzed explanation methodologies for all our results. Enabling us to gain insights into our system and supplying ways to give feedback about the scores.

In addition to that, as our dataset is composed of argumentative essays, whose some topics can be polemical, then the human evaluator can be prone to evaluate the essays according to his beliefs. Therefore, the score may be affected by the experiences of the human evaluator. Such influence can benefit a student or even harm the score, thus influencing the life of a person. To tackle this problem, we proposed a methodology to audit human scores and to measure the degree of bias in texts. To the best of our knowledge, this is the first proposal to deal with human bias. Still, it is imperative to diminish the harm of such a problem.

1.2 Main Goals

The main goals of this dissertation are the following: the building of a system to score multi-aspect essays using handcrafted features, and the analysis of bias on human evaluation of essays. The next subsections describe in high level the subgoals and how we accomplished them.

1.2.1 A System for Automatic Evaluation of Multi-aspect Essays

Multi-aspect essays are dissertations that are evaluated not only by a global and general scoring but also by a set of scoring that each one characterizes one aspect of an essay.

The ENEM, for instance, has five different criteria for the evaluation of essays. The detailed description for each criterion can be found in the guidelines for the candidate[2].

Next, we summarize these five criteria.

1. *Mastery of the formal written language.* This aspect demands to respect grammar rules related to spelling, regency, agreement, and semantics. Some examples of grammar rules are vocabulary accuracy and the absence of oral language usage.

2. *Understanding the writing proposal and applying concepts from the various areas of knowledge to develop the theme, within the structural limits of an argumentative essay.* Regarding this aspect, the student needs to prove that understood the theme, organized the ideas, and applied the ideas in an argumentative dissertation. For instance, besides the student's demonstration of knowledge in different areas like biology, cinema, and literature, he or she should coherently relate such information.

3. *Selection, association, organization, and interpretation of information, facts, opinions, and arguments to advocate a point of view.* In this aspect, it is required of the student to be accurate about the facts. For instance, the student could mention statistics or quotations to prove his or her argument.

4. *Demonstration of knowledge of the linguistic elements necessary for the development of the argumentation.* This aspect requires that the student writes a cohesive argumentative dissertation. For instance, in an argumentative essay, there is the main idea that should be connected to the secondary ideas.

5. *A solution proposal for the problem addressed that respects human values and considers sociocultural diversity.* The solution proposed is an aspect that demands that the student exposes his or her suggestion of solution and how to accomplish it. For instance, the student should propose a viable solution for the addressed problem.

These aspects are essay rubrics specifics to ENEM exam. However, they also indicate the different facets and the holistic character of essay evaluation. Considering essays evaluated according to these five aspects and AES systems, we listed the following goals:

1. To build a baseline proposal composed only of general features.

[2] `http://download.inep.gov.br/educacao_basica/enem/downloads/2019/redacao_enem2019_cartilha_participante.pdf`

2. To build a framework proposal composed of specific features such that it is capable of assigning a score to each of the five aspects described.

3. We aim to perform a feature analysis for each aspect. We understand that it is helpful for educational purposes to provide explanations to the given grades.

The methodology we propose to accomplish these goals is to implement an AES framework that is comprised of standard features of AES systems [Attali and Burstein, 2006; Zesch et al., 2015] and, like previous works, this is our baseline system. To perform the second goal, we developed features specifically for each aspect, improving our baseline system. Our feature analysis is performed using ablation tests and SHAP values [Lundberg and Lee, 2017].

1.2.2 An Analysis of Bias on Human Essay Evaluation

Human evaluation of argumentative dissertation can come with what we call "bias on evaluation," which is characterized by the disagreement between evaluator's opinion and the opinion expressed in the essay. There are some proposals to bias detection on written language, like Recasens et al. [2013] that introduced a methodology to detect linguistic cues that distinguish bias on Wikipedia revisions. Also, Søgaard et al. [2014] presented some strategies to handle bias language with NLP methods.

Considering written comments of human evaluators about essays, we present the following goals to this dissertation:

1. To detect the bias quantitatively on comments of human evaluators in essays;

2. To measure the influence of the human bias on essay scoring.

We intend to accomplish these goals through a lexicon list of words that indicate subjectivity.

1.3 Dissertation Statement and Contributions

The automatic correction of multi-aspect essays is a complex problem since each aspect is modeled according to a different set of features; the students must understand the influence of such features in their grades. Another issue that might concern students is how human bias affects their grades, which is an issue that has not been analyzed yet. Therefore, we aim to propose a new methodology that is efficient regarding human bias to score multi-aspect essays, and whose scores are explainable to the student.

The contributions that follow this statement are categorized as empirical and theoretical. In terms of theoretical contributions,

- we proposed a categorization of bias to detect in a text since this is a word with several meanings (Section 4.1);
- we introduced the definition of the degree of subjectivity of a text, and from this, we proposed the definition of biased text (Section 4.1).

In terms of empirical contributions,

- we built open corpora of labeled Portuguese essays (Section 3.1);
- we proposed an Automatic Essay Scoring (AES) framework for the Portuguese language (Section 3.2);
- we proposed a Multi-aspect essay scoring system (Section 3.2);
- we analyzed in detail the relevant features for our Portuguese AES framework for each aspect (Section 3.3.1);
- we introduced a new unsupervised method to detect bias in texts (Section 4.3);
- we proved that biased labeling harms the classification results (Section 4.4).

In addition to theoretical and empirical contributions, we also introduced some thoughts about the process of human essay scoring and education in general. As thoughts,

- we introduced the discussion about the adoption of an essay as the primary selection way to get into a college;
- we proposed a reflection about the influence of the stance of a human evaluator in an essay score;
- we proposed the possibility of the introduction of an AES system in one of the biggest high school exams in the world;
- we also proposed a thought about the impact of such a system in a non-standardized educational country like Brazil.

1.4 Dissertation Overview

This dissertation proposal is organized as follows:

Chapter 2 provides a comprehensive review of related work about AES systems and biased language detection. In addition to AES systems, we also review problems related to automatic essay scoring like argument evaluation and coherence detection.

Chapter 3 describes the methodology we developed to score essays. We detail our features and why we choose them. This explanation is essential since we intend to perform a feature analysis that shows why a different model is necessary for each aspect.

Chapter 4 depicts which linguistic cues we employed for the detection of the biased language in the human evaluator's comment. In addition to that, we provide a qualitative analysis of human bias on essay evaluation and how it influences scoring.

Chapter 5 concludes this dissertation proposal with some thoughts and future directions to this work.

Chapter 2

Background and Related Work

Our work is composed of two lines of work that are connected by the research area. Therefore we divided our related work into three sections. The first section introduces the technical background needed to understand our work. The second section describes AES systems in English and non-English languages, strategies employed by AES systems to score essays, and related problems that are connected to AES systems, like argument scoring. Moreover, the third section describes research on the bias of written language.

2.1 Background

The steps of automatic essay scoring include definitions of how the text can be represented, how the representation can be used mathematically to predict the score, and the metrics to measure the performance of the automated score system. Next, we detail each one of these topics.

2.1.1 Representing a text

In the machine learning process, the entity to be learned, which is called "document", is represented in a way that the algorithm should be able to build a model or a mathematical function based on the document representation. The model or the mathematical function then evaluates new documents and perform some inference about them. Figure 2.1 describes this scheme in a high-level way. In the figure, x is a vector that represents a document that the machine learning algorithm makes a mathematical model upon it. One way to build this vector is through the feature extraction of a text. Examples of features of a text are the number of words, the number of verbs, and the number of

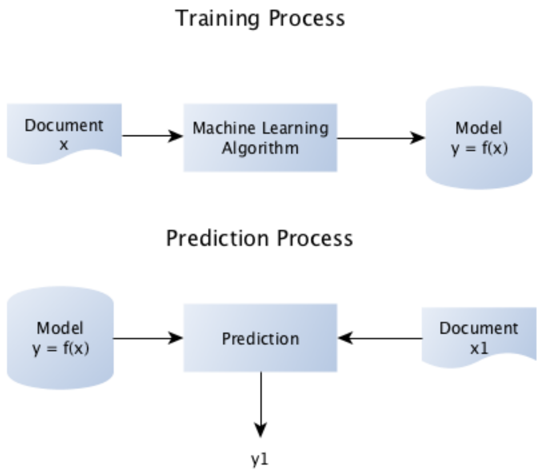


Figure 2.1. The learning process of a document

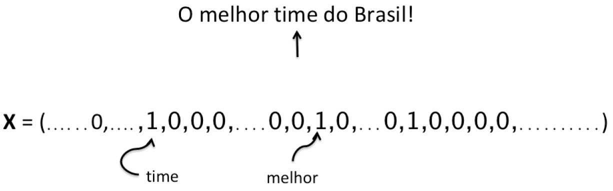

O melhor time do Brasil!
X = (......0,....,1,0,0,0,....0,0,1,0,...0,1,0,0,0,0,..........)
time
melhor

Figure 2.2. An illustration of an one-hot vector representation of a text

exclamation marks. For instance, if a text is represented by $x = \{107, 10, 2\}$, then it owns 107 words, 10 verbs, and 2 exclamation marks.

Another way to build the vector x is by using bag-of-words representation [Manning et al., 2008]. This representation assembles a vocabulary of size N, and then it constructs a vector of dimension N for each document. Each position in the vector represents a word in the vocabulary, and it can hold boolean values like true or false, the word's frequency in the document, or the term frequency-inverse document frequency (tf-idf). The similarity between two documents u and v is computed by employing a distance equation, for instance, cosine formula (Equation 2.1).

$$similarity = \frac{\sum_{i=1}^{N} u_i v_i}{\sqrt{\sum_{i=1}^{N} u_i^2}\sqrt{\sum_{i=1}^{N} v_i^2}} \tag{2.1}$$

Also, x can be represented by a one-hot vector, which is the simplest way to represent a text. Figure 2.2 shows a text and an illustration of its one-hot vector.

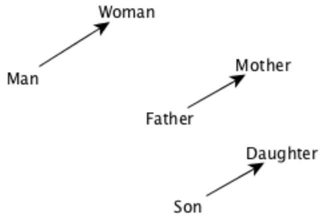


Figure 2.3. The vectors offsets for three word pairs

All the types of representations described until now lack semantic information about the word. However, it is possible to represent such information using a more complex representation called dense vector or word embeddings [Goldberg, 2016]. This representation is built by an algorithm called Neural Network (NN), a type of unstructured learning algorithm in a given dataset.

Algorithms that assemble these representations consider the context of words, unlike bag-of-words. Therefore words applied in similar contexts present similar embeddings. Figure 2.3, which exemplifies the gender relation, shows three-word pairs and the offset between them. Essays also can be represented as embeddings [Alikaniotis et al., 2016; Dong and Zhang, 2016], where each word in the essay are represented by one dense vector.

In the following section, we detail the neural network functioning.

2.1.2 Learning strategies

There are two types of learning strategies used to score essays: structured learning and unstructured learning. In this section, we summarize the definitions of each strategy.

Structured learning is a methodology that employs a set of examples S whose classes C are known beforehand. The goal is to learn from examples and to infer the classes of new inputs. The process of learning from a structured algorithm can result in a set of rules, linear mathematical functions, or any linear mapping. In this dissertation, the learning of the algorithm is performed in the features extracted from the essays.

Linear regression is a structured learning methodology, and we employ it as one of the machine learning techniques to score essays. For instance, suppose that the vector of features of the essays is two dimensional, and we want to predict the score Y, then the linear regression produces a hyperplane equation that minimizes the distance

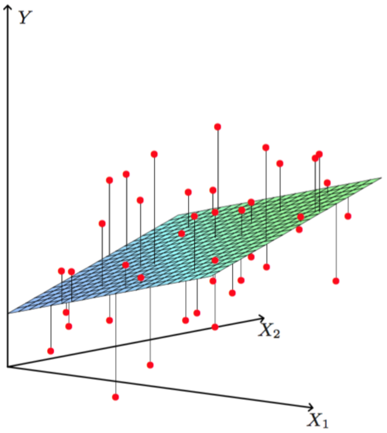


Figure 2.4. Linear regression illustration, from Friedman et al. [2001]

to the training set points. Figure 2.4 illustrates this scenario.

The class of the structured algorithms employed are called regression algorithms, whose input is a set of documents as features vectors, and the output is a real number.

Unstructured learning is a kind of non-linear learning strategy, and the most popular technique is the Neural Network (NN). There are several types of NN, but according to Goldberg [2016] we can think of a NN as a function $NN(x)$ whose input is a vector x which size is N and the output is a vector y which size is L. For this kind of technique, the description of features is unnecessary since it employs optimization methods to search for the best features and the best combination of them [Goodfellow et al., 2016].

A NN is composed of neurons arranged in layers and connected by links. Each link between neurons is associated with a weight matrix, which is initially set up with random values. When a vector passes through this link, the vector is multiplied by a weight matrix, and then a neuron combines this result with the results of other links. Next, the neuron applies an activation function to this combination, and this process continues until the final layer is achieved.

The activation function applied by a neuron is usually a non-linear function that aims to normalize the neuron result and to build a non-linear model for the given dataset. The most used activation functions are:

1. *Hyperbolic tangent*, which normalizes the input to values between -1 and 1.

2. *Logistic*, which normalizes the input to values between 0 and 1.

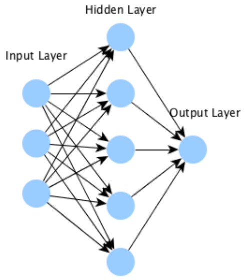


Figure 2.5. An illustration of a neural network

3. *Rectifier linear unit* (ReLU), which normalizes the input to values between 0 and $+\infty$.

The layers in a neural network are classified in the following three types: the input layer, the hidden layer, and the output layer. Figure 2.5 depicts an NN with one input layer that has three neurons, one hidden layer that has five neurons and one output layer that has one neuron.

Once the value in the output layer is computed, the error between the expected output t and the predicted output y is calculated. A standard metric to measure the error E is the L2 norm (Equation 2.2). In a step called *backpropagation*, the error is passed on to the earlier layers, and the weight matrices W are updated according to the gradient descendent technique. This technique is based on the derivative operation, which is a calculus operation that can be used to find the local maximum of a function. In this scenario, we want to find the output that results in the minor possible error value; hence, we add a minus to the computed derivative.

$$E = \sum_i \frac{1}{2}(t_i - y_i)^2 \tag{2.2}$$

When we are building the representation of a word, the desired output t is modeled according to the *likelihood* formula (Equation 2.3).

$$L(\mathbf{x}, \overrightarrow{y}; W) = -\sum_k y_k \log p_k \tag{2.3}$$

The intuition of the Equation 2.3 is that it considers the probability p_k given the word k [Koehn, 2009]. The probability is computed based on the dataset we have. Therefore the likelihood tries to calculate a distribution of probability according to the words of a dataset. This distribution computed should be maximized to be as similar as

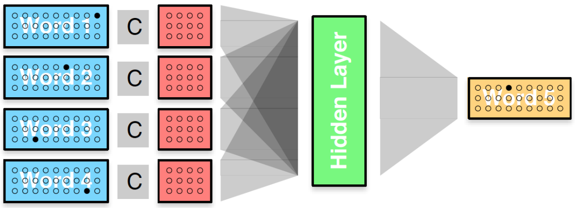


Figure 2.6. An illustration of a neural network to build a Neural Language Model (Figure from Koehn [2009])

possible to the real word's distribution. The result of this process is a *neural language model*.

In a neural network that builds the language model, the input layer presents a specificity compared to a generic neural network. In the input layer, the same weight matrix C is shared among the input neurons. Figure 2.6 depicts such a specific network. The idea is to explore the hypothesis that the words that appear in the same context probably appear together in other pieces of text, then, intuitively, they share similar weight matrix as well.

2.1.3 Metrics

The performance of AES systems is usually evaluated through the Quadratic Weighted Kappa (κ) metric [Brenner and Kliebsch, 1996]. Other metrics employed are Spearman's rank correlation coefficient, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

The κ measures the agreement between two evaluators, and it varies between 0 (none agreement) and 1 (full agreement). In this proposal, one evaluator is the AES, and the other is the human evaluator. Thus, assuming a set of essays E that is scored according to N possible ratings by two evaluators, one human H, and one automated A, then an N-by-N histogram matrix O is built based on the ratings of the essays. Each element $O_{i,j} \in O$ is the number of essays that received a rating i assigned by evaluator H and a rating j assigned by evaluator A. Next, an N-by-N matrix of weights W is computed according to Equation 2.4. Then, considering that there is no correlation between rating scores, an N-by-N histogram matrix of expected rating E is computed through the outer product between the vector of predict scores and the vector of human evaluator scores. The matrix E is normalized such that E and O have the same sum. The κ formula is fully described by Equation 2.5.

$$W_{i,j} = \frac{(i-j)^2}{(N-1)^2} \tag{2.4}$$

$$\kappa = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}} \tag{2.5}$$

Spearman's correlation ρ is a metric of rank correlation, and it evaluates the relationship between two variables A and H of size n [Dodge, 2008]. This correlation assumes that there is an order in the sequence of values of A and H. Then, the two sequences are sorted in increasing order, resulting in the two sequences R_A and R_H. Spearman's correlation formula is described in Equation 2.6.

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} (R_{A_i} - R_{H_i})^2}{n(n^2-1)} \tag{2.6}$$

, where R_{A_i} and R_{H_i} are respectively the i-th element of R_{A_i} and the i-th element of R_{H_i}.

A less used metric is the Root Mean Square Error (RMSE), which measures the error between the predicted score and the human evaluator score (Equation 2.7). RMSE is less used to assess AES performance because it is usual human evaluators score essays differently. Although it is possible to employ RMSE to evaluate adjacent scores, it is hard to determine what are the best adjacent scores.

$$RMSE(A,H) = \sqrt{\frac{\sum_{i=1}^{n} (A_i - H_i)^2}{n}} \tag{2.7}$$

Finally, a more intuitive metric to the human is the Mean Absolute Error (MAE), which is the absolute difference between two continuous variables, and the average between these values. The Equation 2.8 describes the MAE formula.

$$MAE(A,H) = \frac{\sum_{i=1}^{n} |A_i - H_i|}{n} \tag{2.8}$$

2.2 Automatic Essays Scoring Systems

In the '60s, the Project Essay Grade (PEG) [Page, 1994] was proposed as the first Automatic Essay Scoring (AES) system. Since then, several proposals of AES systems have been developed, mainly for the English language. Thus, the first section presents a review of structured AES systems. The second section describes the main unstructured strategies employed by AES systems to score essays. Finally, the third section lists some NLP tasks that are related to essay evaluation.

2.2.1 Structured AES systems

There are several modern AES systems, and most of them employ a structured approach to score essays. The usual methodology is to propose some handcrafted features and then apply a regression algorithm to build a model. One of the first commercial AES, e-rater [Attali and Burstein, 2006], used this methodology. The e-rater employs a set of ten features that is related to meaningful writing skills [Burstein et al., 2003]. These features are divided into the following categories: 1) grammar, usage, mechanics, and style measures; 2) organization and development; 3) lexical complexity; 4) prompt-specific vocabulary usage. The first category is composed of the division of the following errors by the number of tokens: grammar, usage, mechanics, and style measures. The second category has two features that are related to discourse elements, like introduction and conclusion. Such features are the number of missing discourse elements and the average length of discourse elements. The third category comprises the following two features: vocabulary level [Breland et al., 1994] and the average word length in characters. In the fourth category, there are two features: which of the existing six score values the essay is most similar regarding the vocabulary usage, and how similar is the essay regarding essays with the maximum score. This set of features is then used in a weighted linear transformation to score an essay. To evaluate e-rater, Attali and Burstein employed a set of 25,000 English essays about 64 topics. The results obtained with this simple scheme are robust and are similar to human scoring.

Other AES systems [Chen and He, 2013; Zesch et al., 2015; Liu et al., 2018] use similar strategies; they employed handcrafted features and classical machine learning algorithms to score an essay. The baseline of such systems is like the implementation of Attali and Burstein [2006], and then they propose more sophisticated features to improve the scoring prediction. Such a baseline is frequently employed due to the complexity of building an entire system to predict the score of an essay, which comprises other NLP tasks.

Although handcrafted features can provide explanations understandable for humans, they can be challenging to develop and to implement. Languages other than English are even harder to develop due to the lack of resources. Zesch et al. [2015] evaluate their strategy in an English dataset and a German dataset, and the results for the English language are far more superior than that for the German. The authors' attributed these results to the low accuracy and the lack of diversity of non-English NLP tools. Other AES non-English systems have already been proposed with similar results. Like Kakkonen and Sutinen [2004], which proposed an AES system based on Latent Semantic Analysis (LSA) to evaluate Finnish essays of undergraduate students.

Östling et al. [2013] described an AES system for the National High School Exam of Sweden. This system comprises classic features of AES systems and specific features of the Swiss language. The results showed that the Swiss system is non-competitive for commercial use, but it presented promising results. In addition to these systems, others evaluate essays of non-English language as a second language, like Estonian [Vajjala and Loo, 2013], Chinese [Changhuo et al., 2015], and German [Hancke and Meurers, 2013].

Some essays also present rubrics, and also it is challenging to handcraft features for each rubric. As far as we know, only Napoles and Callison-Burch [2015] proposed a strategy using handcrafted features and structured learning. One task proposed by the authors is the prediction of each rubric, therefore similar to our task. To score each rubric, the authors employed a linear regression and represented each essay with complexity features. Some of these features are the average number of words in characters, the number of tokens, the number of sentences, the number of discourse markers, the Flesch-Kincaide grade level, the number of a proper noun, the number of clauses, the ratio of each type of POS tag. The total number of features implemented by the authors was at least 52. As several human evaluators assessed the essays, the authors proposed a strategy that the variance between human scoring is diminished by a statistical model of grades of each teacher. The AES system exhibited a robust performance, and the authors were successful in the demonstration of the variance between teachers scoring criteria. Probably the robust results are due to the numerous labels for each essay, which is not a very likely scenario in the real world.

One of the main goals of this dissertation is also to build an AES system; however, our system deals with issues that are different from the previous works. Firstly, each aspect or rubric of our task is unique. Therefore we need to assess if the features proposed by other works are suitable for our aspects. Secondly, our dataset is in the Portuguese language, which lacks robust tools for NLP tasks. Thus it is a big challenge to build robust features to learn. Thirdly, the type of essays we handle, like the ENEM exam, allows subjective evaluations. Usually, the topics are polemical, and the government guidelines to the evaluator are not objective [Mendes, 2013].

2.2.2 Unstructured AES system

Unstructured learning techniques are trendy nowadays, and they are state-of-the-art for many NLP tasks, including automatic essay scoring. Alikaniotis et al. [2016] proposed an automatic scoring method based on the Long-short Term Memory (LSTM) network, which is a type of Neural Network. Then, the authors extended the network to include a linear regression strategy that assigns a weight for each embedding of the essay. This technique achieved robust results, and the authors explained the scoring through the weights that most contributed to the final scoring in the linear regression.

Dong and Zhang [2016] is another work that employed a type of neural network called a convolutional neural network (CNN). Like other works, basic handcrafted features and regression algorithms were chosen as the baseline. For the proposed strategy, dense vectors [Mikolov et al., 2013b] were employed to represent the essays, and then a two-layer CNN was applied. The results of the proposed approach were superior compared to the baseline.

Recently, other AES strategies that use unstructured learning were proposed. Like Dong et al. [2017] that applied CNN and LSTM to score essays, but differently from the previous works, the authors also employed an LSTM to build the representation of the text. The explanation to employ this kind of representation is that LSTM is better to learn long sequences, and then features like coherence and cohesion are captured by the vector representation. Coherence is an essential issue in the scoring task, then Tay et al. [2018] also proposed a methodology to consider this feature. The approach is a more complex architecture of a neural network, and it captures coherence in an end-to-end fashion as well. Both works show that considering coherence leverages the results for automatic essay scoring. In addition to coherence, Cozma et al. [2018] proposed to apply string kernels to improve state of the art result in the AES task. The results achieved are superior to the previous works.

Although these works that employ neural networks presented robust performance, they still leave unexplained some aspects that feature engineering is helpful. Given the nature of the automatic essay scoring task, explanations are not only desirable but necessary. In addition to that, unstructured learning methods are more suitable for essays in the same domain, and as they usually need a massive amount of data is not a practical solution in some scenarios. Our dataset, for instance, is composed of several topics, each one with few essays. Therefore unstructured learning using embedding representations is not suitable for our datasets.

2.2.3 NLP tasks related to Automatic Essay Evaluation

The evaluation of essays is a complex problem, then its division into smaller problems is a natural thing to do. For instance, argumentation is a writing skill that the student develops through practice and the teacher's feedback. Hence, several works tackle problems specifics to some writing skills, like grammar, coherence, cohesion, prompt adherence, and argumentation mining. The development of these tasks helps the evolution of AES systems; therefore, in this section, we briefly present them.

Grammar error correction (GEC) was the goal of two shared tasks in the Conference on Natural Language Learning (CoNLL) [Ng et al., 2013, 2014], since then several proposals to improve the results of this task have appeared. Dale and Kilgarriff [2011] described a dataset for the GEC task and the strategies of six teams to resolve the task. Other competition [Bryant et al., 2019] proposed a GEC task that includes grammatical, lexical, and orthographical errors. The results of all these competitions indicated that GEC is a complex task to solve.

The state-of-the-art strategies for the GEC task use deep learning. Yuan et al. [2016] detailed a strategy to correct grammar errors employing statistical machine translation (SMT), which translates an incorrect sentence to a correct one. In this strategy, the SMT system builds a list of candidates to the translation; then, another algorithm is employed to rank the candidates. To propose a new metric to assess the performance of systems that correct grammar errors, Bryant and Ng [2015] provided the dataset from Ng et al. [2014] for ten human experts to annotate grammar errors. One of the conclusions of this work is that as one human versus nine humans achieved a maximum of 70% f-measure, then it is unlikely that the machine reaches a performance beyond that.

Coherence, cohesion, and prompt adherence are tasks addressed by some works that aim to score essays. The definition of each of these tasks is broad; nonetheless, most of the computational approaches narrow their definitions.

In the coherence task, most of the approaches are based on the Rhetorical Structure Theory (RST) described in Manning et al. [2008], and its authors state that "A text is coherent when it can be explained what role each clause plays concerning the whole." Based on this definition, the coherence task aims to measure how coherent a text is. The result is usually a score that was learned by applying a machine learning algorithm in texts scored by humans. This procedure was followed by Barzilay and Lapata [2008], which computationally modeled the coherence definition as the degree of relatedness between two consecutive sentences. To measure the degree, the authors first build a matrix of entities. In this matrix, the columns are entities, which means people, organizations, among others, and the lines are sentences. Each entity is an object, a subject, or neither one in a sentence. Hence, dense columns mean coherent texts, because the entities appear in consecutive sentences, thereby preserving the text's coherence. Finally, this matrix is used to evaluate the coherence of two tasks: sentence ordering and summary coherence rating. Lin et al. [2011] employed a similar definition of coherence; however, the arrangement of discourse relations is used instead of entity relations. They encoded the features of such relations to evaluate the coherence of texts, and then return a score. The state-of-the-art of the coherence task, though, was achieved by neural network approaches, which relies more on computational modeling than linguistic modeling [Li and Hovy, 2014; Li and Jurafsky, 2016].

Concerning cohesion, the most common strategy is to identify the discourse elements that contribute to the text's organization. The types of discourse elements are the introduction, the prompt, the thesis, the main idea, the supporting idea, the conclusion, and others [Burstein et al., 2001]. Song et al. [2015] used a supervised machine learning algorithm to determine a chain of sentences that are locally or globally cohesive. A chain of sentences is globally cohesive when they are part of the same discourse element, but they are separated. Other works that approached cohesion are Persing et al. [2010] and Burstein et al. [2001].

Prompt adherence is the task that evaluates if the essay is related to the topic described by the given prompt. Few works presented solutions to this task, Higgins et al. [2004] proposed a strategy that assesses each sentence if it is good or lousy adherence to the prompt, while Persing and Ng [2014] proposed a model that scores the whole essay regarding the prompt adherence.

Argumentation mining is a task addressed by several works in the NLP field, due to the wide range of applications that arguments are related. Some examples of applications that require assessment of argument's quality are online debates [Habernal and Gurevych, 2016; Lukin et al., 2017] and essay evaluation [Madnani et al., 2012; Stab and Gurevych, 2014; Persing and Ng, 2015; Stab and Gurevych, 2016; Ke et al., 2018].

The structural elements of an argument are composed by a claim and one or more premises. The claim is connected to a premise by a support or attack relation. The identification of argumentative elements and non-argumentative elements, the classification of argumentative elements into claim and premises, and the identification of relation type between claim and premises are subtasks of argumentation mining. Stab and Gurevych [2014] provides the following example of arguments containing a claim in boldface, and a premise underlined.

Example 1. **It is more convenient to learn about historical or art items online**. With the Internet, people do not need to travel long distance to have a real look at a painting or a sculpture, which probably takes a lot oftime and travel fees.

The cornerstone of an argument is the claim. In the example above, the writer asserts that it is easier to learn online about historical or art items. The premise should support the claim by giving it grounds to persuade the reader. To detect each of these components, humans label the components of the arguments in the text following some guidelines. Next, a machine learning strategy is applied in the labeled data to build a model that identifies arguments in unlabeled data. The detection of the components of the argument is an important task that helps to leverage the evaluation of the argument strength that can be represented by a score.

Most of the work done in argumentation mining focus only on the detection of the argument components, like Madnani et al. [2012], that proposed three methodologies to classify argumentative elements into claim and premises. The best performance was achieved by the methodology that employed a supervised CRF algorithm combined with rules and prediction at the word level. The results were satisfactory; still, the methodology's performance was weak in the political debates domain. Another proposal is presented by Stab and Gurevych [2014] that besides premises and claims identification, it also recognizes relations between premises and a claim. The strategy is composed of two steps, in the first step, the argument's components are identified, and in the second step, relations between claims and premises are classified between support or attack. To accomplish these subtasks, the authors extracted several features, and then an SVM algorithm is applied to the learning process. A proposal that encompasses all the argument mining subtasks is described in Stab and Gurevych [2016]. As this work identifies all the structure of an argument, it is suitable for writing support systems and for assessing the quality of an argument. Regarding the essays, Persing and Ng [2015] described a strategy that employs a rich set of features and regression to predict the score of the argument strength in essays. The prediction starts with the identification of argument elements, and then the features are extracted from argument elements. The results were promising, and the conclusion was that argument strength is possible to be assessed automatically.

2.3 Biased Language

Bias is a word with a broad sense, and maybe that is why several works approach bias detection from different perspectives. Then we categorized the bias detection task according to the following types:

1. **Detection of explicit bias**. The detection of such bias is conventional in social networks where people want to be clear about their stances.
2. **Detection of implicit bias**. In this situation, someone is unaware of his or her own bias. For instance, when most people think about a nurse, they think a woman nurse and not a male nurse.

Explicit bias is the subject of Yano et al. [2010] work, where American political blogs were employed to analyze the tokens that are relevant to explicit bias. The mechanical turkey labeled sentences from blog posts that the authors collected. The authors analyzed the sentences and concluded that some tokens are biased against the two main American political parties, the Democrats and the Republicans. Iyyer et al. [2014] also aimed to detect explicit bias, but by employing a *Recursive Neural Network* (RNN). The authors achieved superior results compared to previous works. Also, Faulkner [2014] proposed a method to detect explicit bias; however, the author applied classical machine learning techniques in argumentative essays. Regarding explicit bias, many other works were proposed [Zhou and Cristea, 2016; Zafar et al., 2016].

Regarding implicit bias, Gentzkow and Shapiro [2010] proposed a methodology to detect the political bias in newspapers. Their methodology classifies lexicons according to democratic or republican type. The authors employed a collection of speech from Democrats and Republicans politicians, and the most common lexicons in each party were classified accordingly. By using the lexicon lists, the proposed methodology was able to identify the political bias in the newspapers. With a more general domain, Recasens et al. [2013] presented a strategy to detect bias in the Wikipedia articles. The idea of this work is based on the *neutral point of view* (NPOV) policy from Wikipedia, which requests that the editors write an unbiased, impartial, and reliable text [Wikipedia, 2017]. Therefore, the authors proposed a strategy that the main goal is to identify bias introduced by editors and removed by seniors editors. These editions were collected and labeled as biased or not through Mechanical turkey. The features extracted from the sentences were: factive verbs, implications, and other entailments, hedges, and subjective intensifiers. Then, a Logistic Regression algorithm was used to detect the bias of the sentences. The accuracy of the methodology and the humans were different by at most 3%.

Additionally, Søgaard et al. [2014] discuss the bias when humans label data, which is called bias on the ground truth. Then the authors proposed some solutions like semi-supervised learning or the mean between multiple labels. However, semi-supervised learning can worst the problem if the dataset is too complex, and multiple labelers are not always available. We also aim to detect and to analyze implicit bias; nonetheless, we want to perform in a whole different domain, which is the comment's reviewer of argumentative essays. Moreover, we introduce a new unsupervised solution to detect implicit bias. None of the works above proposed a similar method.

Chapter 3

Essay Scoring

We implemented an Automatic Essay Scoring (AES) system employing two types of learning strategies, structured and unstructured. Both types were already explored in other works. However, different from the previous approaches, we predicted five aspects of an essay. Besides comparing the strategies, we analyzed the effect of hand-crafted features in each aspect, revealing the behavior of our structured algorithms and interesting explanations for educational studies. In particular, our experiments aim to answer the following research questions:

RQ1: Which are the most relevant features to a structured technique like regression? Can we have some insight into human evaluation regarding the aspects of essays?

RQ2: How scores are distributed across the essays? How aligned with human raters are the different AES models?

RQ3: Do the Structured and the Unstructured learning present comparable performance in the multi-aspect Automatic Essay Scoring task?

Next, we describe two Portuguese datasets of essays that we built, structured learning and unstructured learning techniques, the experiments performed, the results achieved, and finally, a discussion about the results.

3.1 Corpus

We built two corpora of Portuguese essays. The first corpus is composed of essays that were crawled from UOL Essay Database website[1], and the second corpus is composed

[1] `http://educacao.uol.com.br/bancoderedacoes`

Dataset	#essays	Avg. #tokens	Max. #tokens	Min. #tokens
UOL	2,431	285.19	892	101
Brasil Escola	8,553	323.70	800	100

Table 3.1. Datasets of Portuguese Essays

of essays crawled of the Brasil Escola website[2]. Table 3.1 shows some statistics about corpora.

In both corpora, each essay is evaluated according to five aspects. The ENEM exam also employs the same evaluation aspects. The full description of each aspect can be downloaded on the Brazilian government website[3].

The aspects are the following:

Formal language: Mastering of the formal Portuguese language.

Understanding the task: Understanding of essay prompt and application of concepts from different knowledge fields, to develop the theme in an argumentative dissertation format.

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.

Each aspect is scored according to the scale of Table 3.2, and the final score is the sum of all aspects scores. Table 3.3 depicts the average score assigned by humans for each aspect and final grade in our datasets.

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

Figure 3.1 presents the distributions of the scores for each aspect and the final score in the UOL dataset. As we can notice, the scores are imbalanced for all the aspects.

[2] https://vestibular.brasilescola.uol.com.br/banco-de-redacoes/

[3] http://download.inep.gov.br/educacao_basica/enem/downloads/2019/redacao_enem2019_cartilha_participante.pdf

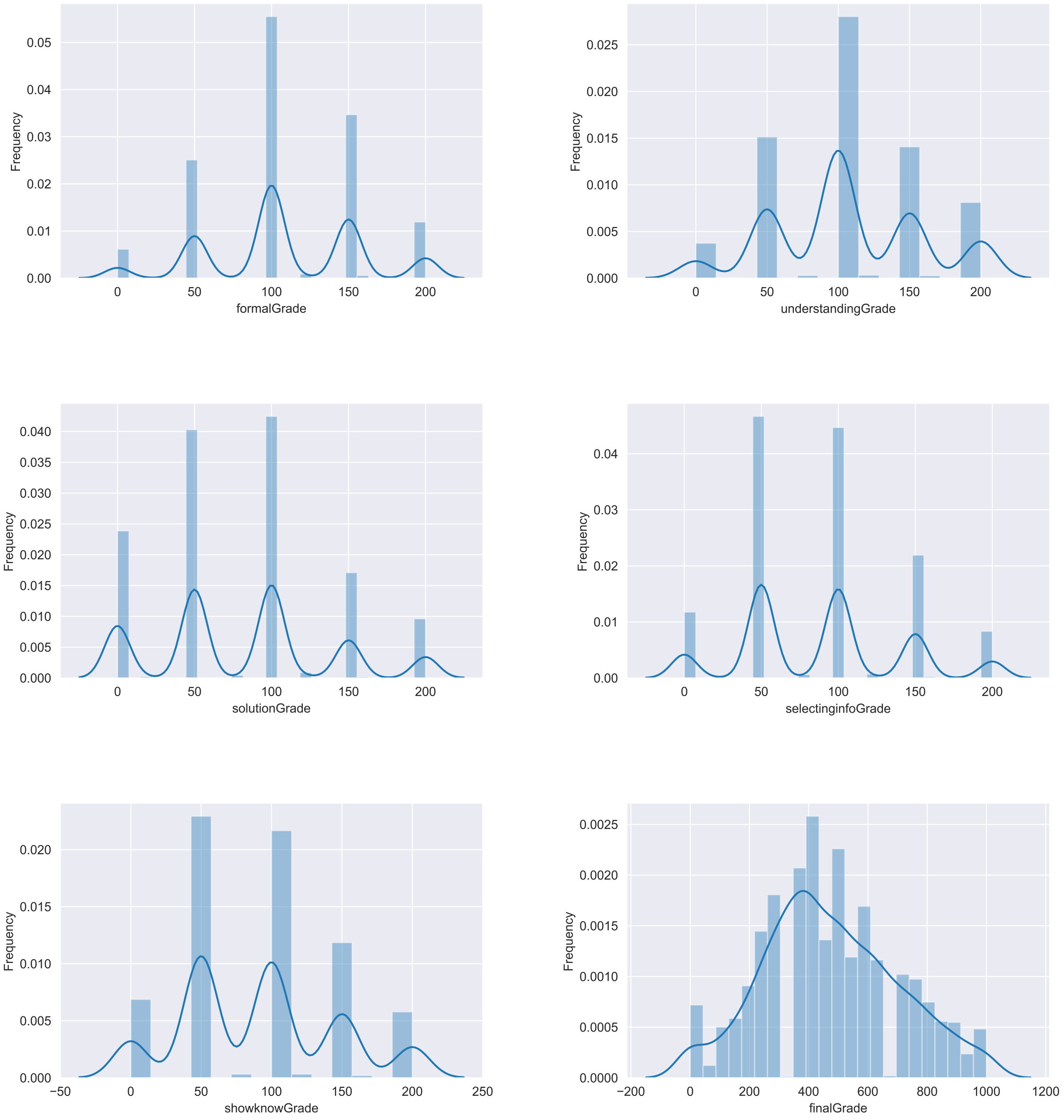


Figure 3.1. Distribution of grades in the UOL dataset for each aspect and final grade

Score	Level
200.0	Satisfactory
150.0	Good
100.0	Regular
50.0	Weak
0.0	Unsatisfying

Table 3.2. Score and corresponding levels

Aspect	**UOL**	**Brasil Escola**
Formal Language	108.12	115.75
Understanding the task	105.65	122.27
Organization of information	88.33	132.13
Knowing argumentation	90.63	131.98
Solution proposal	80.85	145.59
Final grade	473.33	647.73

Table 3.3. Average Score for each aspect and final grade in UOL and Brasil Escola datasets

The distributions of the scores for each aspect in the Brasil Escola corpus is depicted in Figure 3.2. Like the UOL dataset, the scores are unbalanced; however, the scores in the Brasil Escola dataset are higher than the UOL's scores.

3.2 The Handcrafted Features and the learning techniques

The features we elaborated are divided into two main types, domain features that are related to each aspect, Portuguese Language or ENEM test, and baseline features that are based on Attali and Burstein [2006] research. Attali and Burstein work categorized the features into four types:

1. Grammar, Usage, Mechanics, and Style Measures: Raw counts of errors about grammar, usage, mechanics, and style that are computed by the software Criterion [Burstein et al., 2003] and then divided by the total number of words in the essay. As Criterion is proprietary software, the authors did not reveal how the features are extracted.

2. Organization and Development: The score organization considers a standard five-

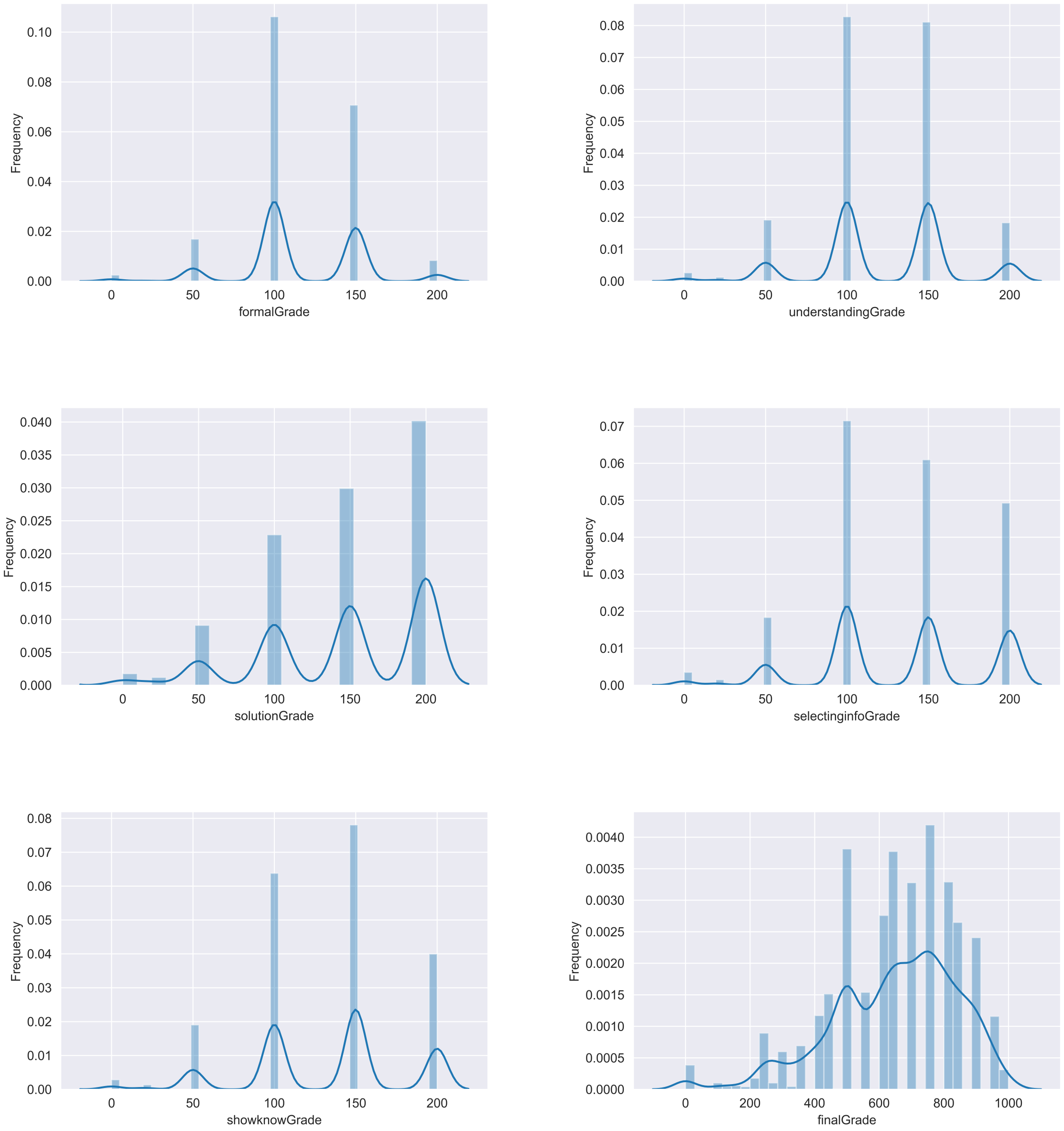


Figure 3.2. Distribution of grades in the Brasil Escola dataset for each aspect and final grade

paragraph essay as the goal and compare it to the input essay. The difference between the discourse elements of the standard and the input is the given score. The discourse elements are manual annotations in the training data that indicate the main idea, thesis, among others. Regarding the development, the score takes into account how much of the discourse elements were augmented (text length).

3. Lexical Complexity: In this category, there are two metrics, the vocabulary level that is computed according to the Breland et al. [1994] definition, and the average word length in characters.

4. Prompt-Specific Vocabulary Usage: Regarding the domain vocabulary, the authors build an array where each dimension is the average coseno distance between essays from a given score and the given input essay. The scores are a discrete set of numbers that are defined according to a scale. Then, if the scale comprises six scores, the prompt feature vector would have six dimensions. The vector that represents an essay is a word frequency vector, and the dimension of the vector is the length of the defined dictionary.

Like many other works, our baseline is based on the features proposed by Attali and Burstein. Also, we proposed some other features that are unrelated to the Attali and Burstein work, but they strengthen our model since they are related to some aspects of our domain problem. Because these additional features are related to our domain problem, i.e., the multi-aspect score prediction of Portuguese essays, we called them domain features.

However, we had to adapt some of the features due to the lack of tools for the Portuguese. Next, we list our baseline features according to the four categories of Attali and Burstein [2006].

1. **Grammar, Usage, Mechanics, and Style Measures**.

   Grammar was checked by CoGrOO [2012], which is a Brazilian add-on to Open Office Writer. According to the documentation of CoGroo, it detects errors like nominal agreement, verbal agreement, and nominal regency. For instance, the agreement between the subject and the verb in the following sentence is wrong and would be detected by the software.

   Example 1. *Nós vai* (We goes).

   For spelling mistakes, we used another Brazilian software[4]. This software finds wrong words based on a dictionary of 600k words and string distance metrics.

[4] `https://github.com/giullianomorroni/JCorretorOrtografico`

Both features were also divided by the number of tokens in an essay; therefore, we employed four features for grammar and spelling errors.

To evaluate style in essays, we applied LanguageTool[5] rules for Portuguese, but also we added some rules suggested by a Portuguese manual of writing [Martins, 2000]. The rules comprise mistakes like wrong preposition use, punctuation errors, and redundancy. The following sentence is an example that is discovered by LanguageTool.

Example 2. *Ele chegou a conclusão final* (He reached the final conclusion).

We employed the number of style errors and the number of style errors per sentence as features. To more details about the rules employed, we recommend the reader to Appendix A.

2. **Organization and Development**.

   There are no tools to evaluate organization and development in the Portuguese language; then, we collected discourse markers in Portuguese grammar [Jubran and Koch, 2006]. Discourse markers are linguistic units that establish connections between sentences to build coherent and knit discourse. Some instances of discourses markers are *logo* (next) and *no entanto* (however). We employed as features the number of discourse markers and the number of discourse markers per sentence. The complete list of discourse markers is open and is available for download[6].

   In addition to that, we count the sentences longer than 70 characters, which, according to Martins [2000] are long sentences.

3. **Lexical Complexity**.

   To evaluate lexical complexity, we used four features. The first feature is the Portuguese version of the Flesh score [Martins et al., 1996], which is a formula that assesses a text and assigns a score that represents the appropriate educational level of the writer. The second feature is the average word length, in which the length is the number of syllables. The third feature is the number of tokens in an essay; the fourth feature is the number of different words in an essay.

4. **Prompt-Specific Vocabulary Usage**.

[5] `http://wiki.languagetool.org/java-api`

[6] `https://github.com/evelinamorim/aes-pt/blob/master/discoursemarkers.txt`

It is desirable to employ concepts from the prompt in the essay. Hence for each essay, we computed the cosine similarity between the prompt and the essay. Both texts are transformed into dense vectors using the following steps. First, for each word in a text document, i.e., a prompt or an essay, we transform it in a word embedding, and then we compute the average of the word embeddings of all words in the document. Finally, we compute the coseno distance between the two vectors. We decided on this strategy since, unlike other works, our datasets comprise many different topics, each with few essays. Then, we think that to build a vocabulary for the domain of each topic is not helpful.

Our domain features are related to ENEM skills. These features distinguish our problem since ENEM skills present the challenge of score multiple skills. Only recently, there are some proposals to score Portuguese essays, and none of these works addressed ENEM skills or Portuguese language features. Next, we describe our features according to the skill that is related to it.

1. **Formal language**.

   In this category, there are features intrinsic to the Portuguese language, and that can reveal a good or bad use of the language.

   i. *Ênclise*: It is a Portuguese language structure that dictates rules for the placement of some kinds of pronouns after the verb. This structure goes together with a hyphen, which connects the pronoun and the verb. The following sentence is an example of the use of *ênclise.*

      Example 3. *Espero contar-lhe isto hoje à noite.* (I hope to tell you this tonight.)

      Ênclise is unusual to the Portuguese spoken language of Brazil, and it is a construction related to formal language usage. We employed as features the number of *ênclise* and the number of *ênclise* per number of tokens.

   ii. Oblique Pronouns: The use of oblique pronouns in several languages is innately anaphoric; however, their role in a clause is an object that is restored in the phrase. As the application of such elements requires a strong knowledge of syntax, we tried to detect the usage of oblique pronouns in the essays. We use as feature the number of oblique pronouns.

2. **Knowing Argumentation**

i. Indetermination Instruments: According to Perini [2017] the indetermination in Portuguese is the capability to understand the reference to a phrase. For instance:

   Example 4. *essa mesa redonda* (this round table)

   The phrase above is determined because it provides resources to the reader to detect who is being referenced. However, the noun *mesa* (table), without any articles, is less determined than the example above. In the Portuguese language, there are degrees of determination or indetermination, and as stated in Perini [2017], Portuguese is a language abundant in indetermination instruments. The indetermination is a way to make implicit a personal opinion and thus build a more formal text. Thus, considering all the kinds of these instruments, we tried to capture the pronoun "se" postponed, the use of passive, the lack of use of the first person singular, and the use of the first person plural. The first two features are specific to all the romance languages, including Portuguese. Nonetheless, the last three features are common two all languages, including the English. Since the Portuguese language present several tools to these features, and the ENEM exam requires formal language, we decided to categorized indetermination instruments as a domain feature. Then, in this feature, we counted the number of indetermination instruments.

ii. Subordinate Clause: The employment of subordinate clauses demonstrates a higher syntax complexity of a Portuguese text. Hence we built a list of conjunctions used in the construction of subordinate clauses. From this list, we count the number of subordinate clauses in the essay. The list of conjunctions can be analyzed in Appendix B.

iii. Anaphor pronouns: The anaphor elements in the Portuguese language enable to refer terms previously mentioned, thus allowing the construction of logical reasoning in a text of argumentative structure. In addition to that, the anaphor relations reveal the argumentative structure of a text. Considering the relevance of anaphor pronouns in the argumentation process, we include the number of anaphor pronouns as one feature in our system. Also, demonstrative pronouns have anaphoric nature, so we also count their occurrences.

3. **Organization of Information.**

   i. The similarity between sentences: The skill Organization of Information requires text coherence. To measure the coherence in an essay, we computed the similarity between the sentences of the essay. In this process, first, we employ word embeddings trained with the word2vec algorithm [Mikolov et al., 2013b] in the Portuguese Wikipedia. Each sentence is then represented as the concatenation of its words, and the maximum number of words that we consider in a sentence is seven. Next, we compute the coseno distance between a sentence and the sentence right after. We consider as features the average distance and the largest distance of all distances.

   ii. Part-of-Speech elements: Some parts of speech elements are relevant to the cohesion of a text, as stated by Koch [1999]. The adverbs can refer to a verb phrase, the articles precede nouns that refer to an antecedent or subsequent information, the personal pronouns of the third person can refer to nouns, and the conjunctions link elements or clauses in a text. Thus, we build four features based on the proportion of adverbs, articles, pronouns, and conjunctions in an essay.

4. **Solution Proposal.**

   i. Racist Terms: Silva Neto et al. [2017] built in their course conclusion monograph a list of racist terms for the Portuguese language, and we employed this list to count the terms that can infringe human rights. The solution proposal skill requires that the student stands in favor of human rights; otherwise, the essay is scored as zero.

   ii. Conclusion markers: Koch [1999] lists conclusion markers that start a conclusion statement that links two parts of the text. Then, from these words, we employed a thesaurus dictionary[7] and high score essays to expand the list. Some examples are *portanto* (therefore) and *pois* (because). The complete list can be checked in Appendix C.

Table 3.4 summarizes all the features extracted from the essays. The baseline features are based on the work of Attali and Burstein [2006], and as many other AES strategies, we employed their features as a baseline system. The domain features are related to the aspects of the ENEM exam, and our system comprised both types of features.

Structure and unstructured are the main machine learning techniques applied to the Automatic Essay Scoring problem. Due to this reason, we also applied these

[7] `https://www.sinonimos.com.br/`

Table 3.4. The list of features grouped into domain type and general type. The values of all features are of numerical type.

Group	Feature
Domain	pronouns /#tokens
	articles /#tokens
	adverbs /#tokens
	conjunctions /#tokens
	Indetermination Instrucments
	Subordinate Clause
	Anaphor Pronouns
	Oblique Pronouns (i.e. LexicalComplexity)
	Racist Terms
	#demonstrative pronouns (#tokens)
	Similarity between sentences (Average and Maximum)
	#conclusion markers
	#ênclise
	#ênclise / #tokens
Baseline	#sentences longer than 70 characters
	#grammar errors(/#token)
	#spelling errors(/#token)
	#style errors /(#sentences)
	#discourse markers(/#sentence)
	Flesh score
	Average word length (syllables)
	#tokens
	similarity with prompt
	#different words

techniques to the Multi-aspect Essay Scoring problem. The structured algorithms tested are Support Vector Regression(SVR), Random Forest (RF), Linear Regression (LR), and Xgboost. Each one of these strategies is based on a different mathematical theory; therefore, the results can be different for our datasets. We choose them because, generally, they present robust results for different tasks. For unstructured technique, we choose the Multi-Layer Perceptron (MLP) to test, our input is a feature vector, and the target value for the network is the score assigned by a human to the essay.

3.2.1 Statistics of the features

Before the experiments, it is necessary to analyze the features of our datasets and compare them. Next, we show the histograms of our baseline features and domain features.

3.2.1.1 Baseline Features

In Figure 3.3, group 1 of baseline features, we see that there is no disparity between datasets in most of the histograms. The only features that present differences are SpellingCheck and StyleErrors, which display the largest values for Brasil Escola corpus. Since the normalized versions of these features, SpellingCheckNorm and StyleErrorsNorm, presented similar distributions, then the discrepancies are due to the differences in the average size of essays in each corpus (Table 3.1, Section 3.1).

In the second group of Baseline features, Figure 3.4, again, most of the features are look alike, except for NumberTokens and NumberDiffWords. Also, we can attribute these differences to the length of essays of each corpus, as described in Table 3.1 in Section 3.1. Therefore, comparing the Baseline features, we note that the corpora are similar.

3.2.1.2 Domain Features

The Domain features from group 1 show even fewer differences than all the Baseline features. In Figure 3.5, we can observe that all distributions are look alike.

On the other hand, in the second group of Domain features, Figure 3.6, we observe a divergence in the distributions of Pronouns and ConclusionMarkers. Since ConclusionMarkers are not normalized, then we can guess that higher values in the Brasil Escola dataset are due to its bigger essays. However, the Pronouns feature is normalized by tokens. Thus, if this feature emerges in some model in a dataset, the reason can be related to this discrepancy.

3.3 Experiments and Results

We performed experiments using structured techniques like Support Vector Regression(SVR), Random Forest (RF), Linear Regression (LR), and Xgboost, but also we used unstructured techniques like Multi-Layer Perceptron (MLP). We implemented the different AES models using Scikit-learn[Pedregosa et al., 2011].

Next, we report the results obtained from the execution of the experiments. We employ κ and Mean Absolute Error (MAE) as the evaluation metrics. Besides the ASAP challenge at Kaggle[8], several works employ **quadratic weighted kappa**(κ) as the evaluation metric [Zesch et al., 2015; Chen and He, 2013; Attali and Burstein, 2006], which aims to measure agreement between human evaluation and machine scoring.

[8] https://www.kaggle.com/c/asap-aes/details/evaluation

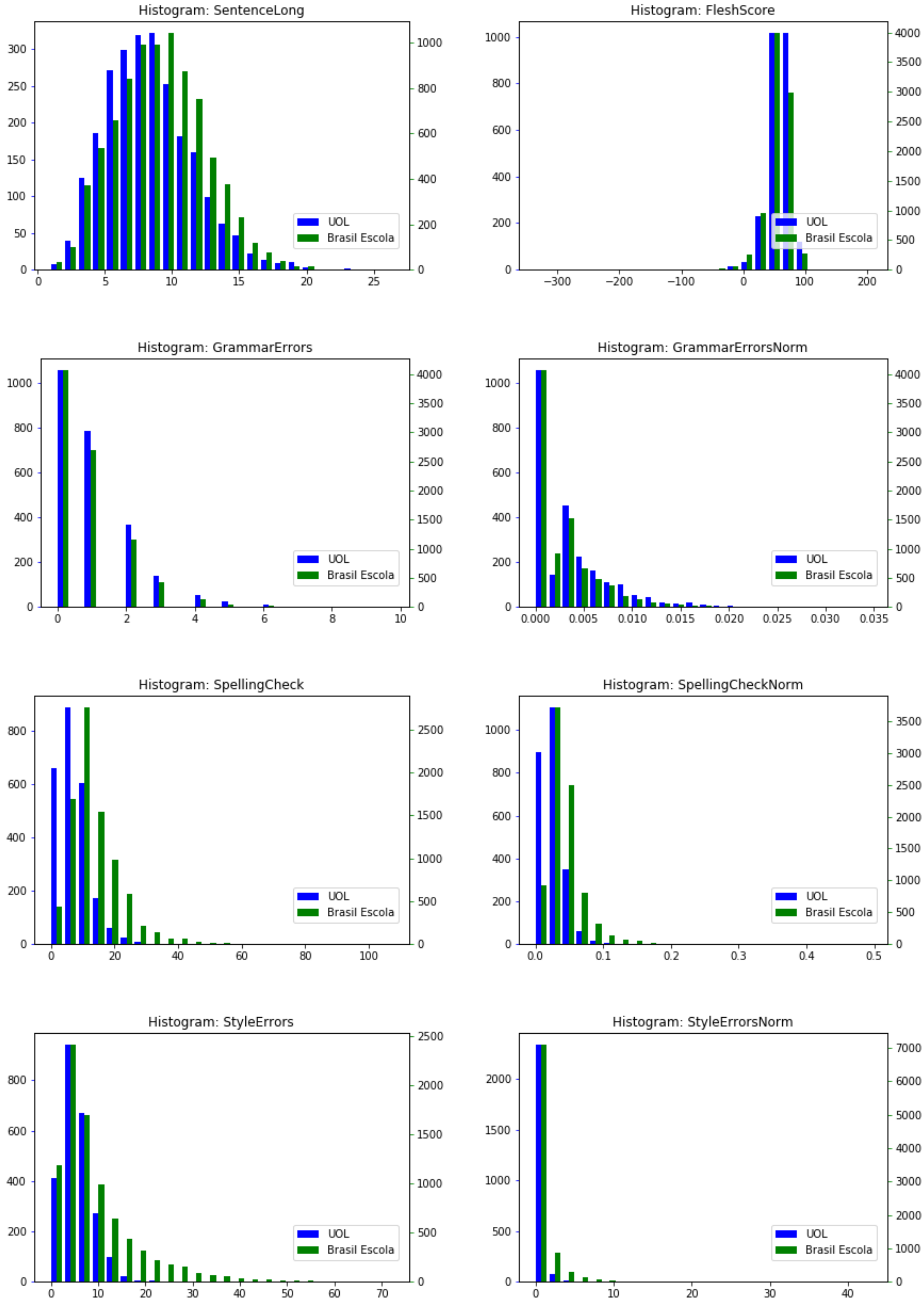


Figure 3.3. Group 1 of Baseline features

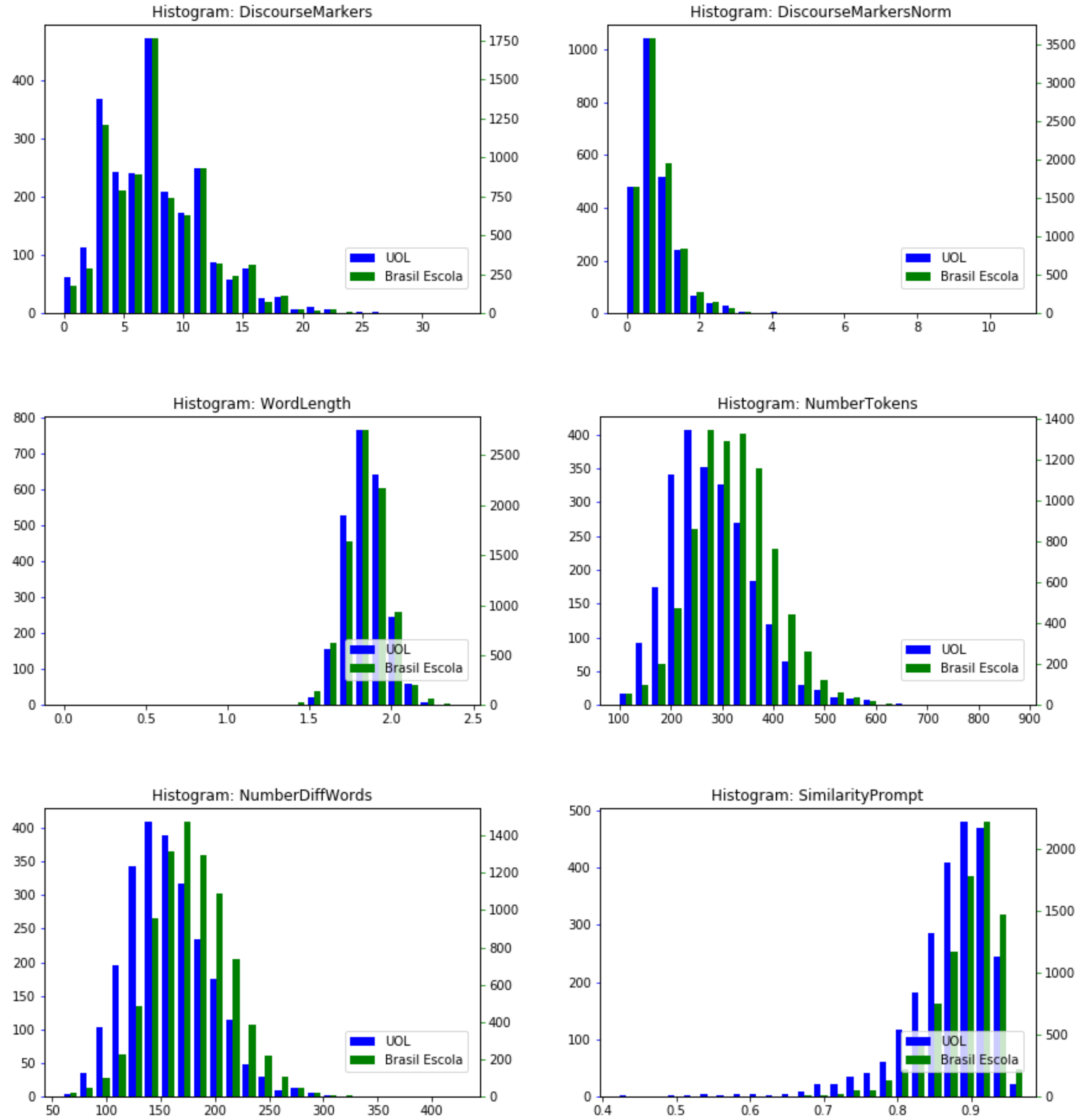


Figure 3.4. Group 2 of Baseline features

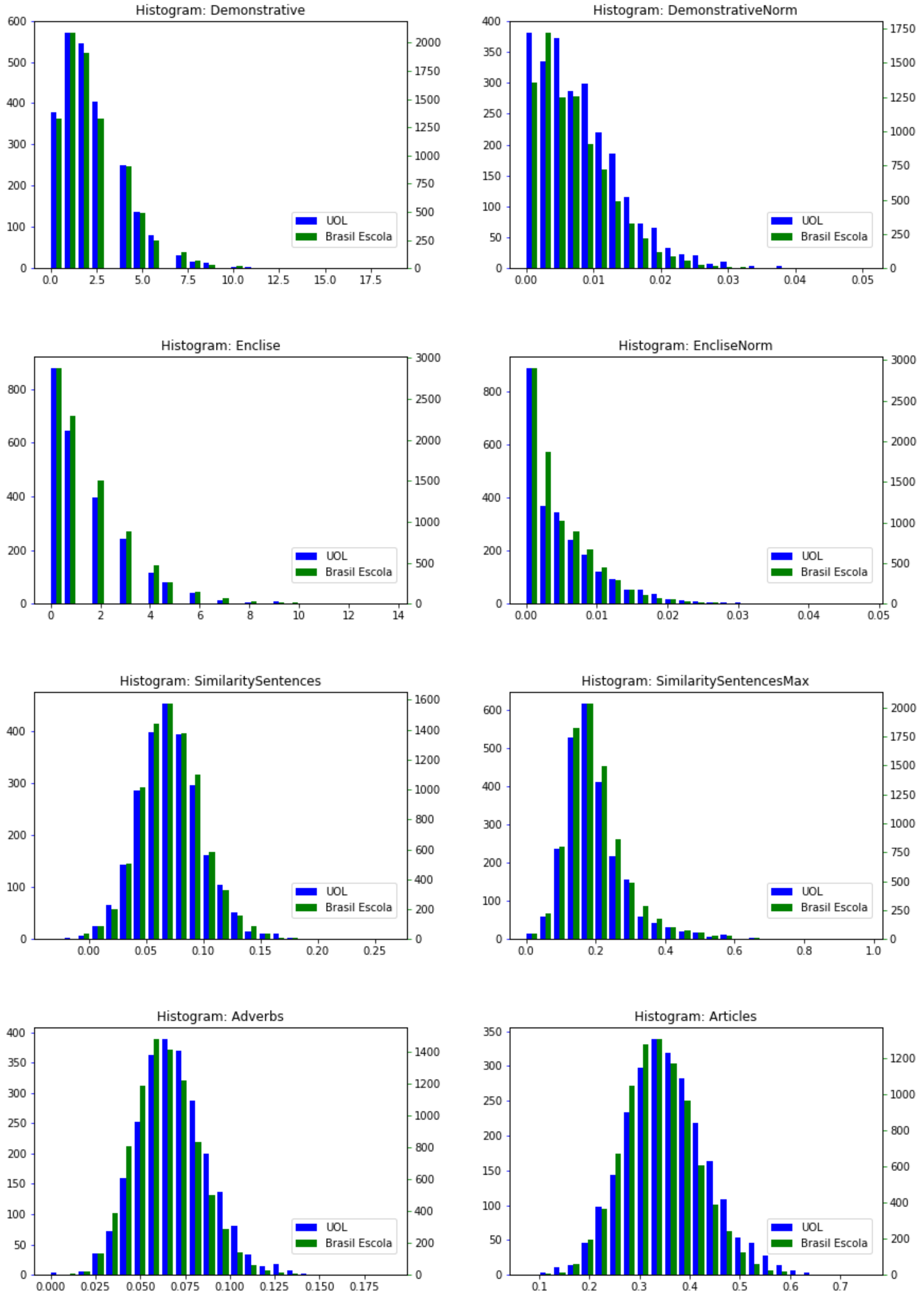


Figure 3.5. Group 1 of Domain features

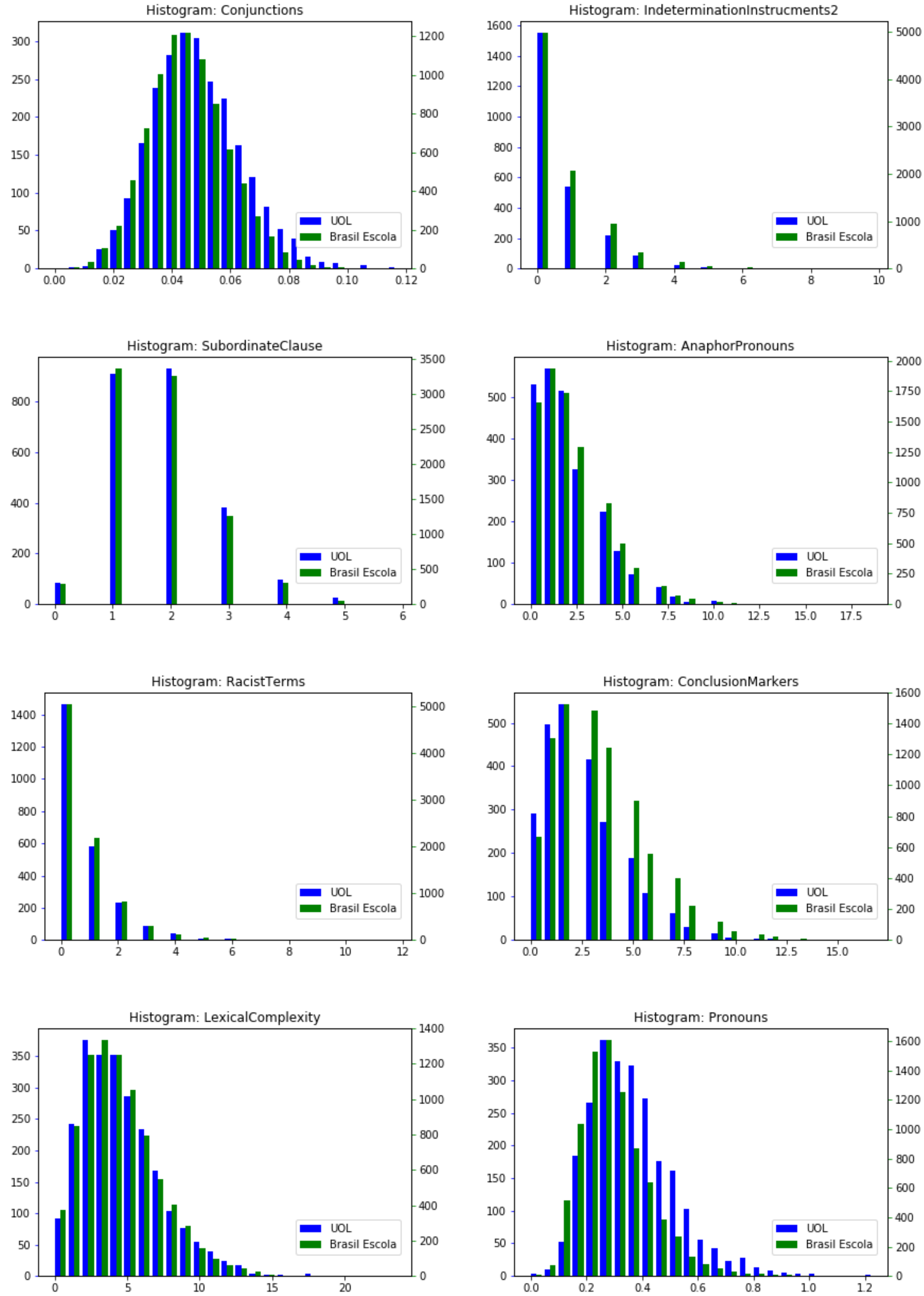


Figure 3.6. Group 2 of Domain features

Table 3.5. Results for each grade aspect for UOL Dataset

Grade Type	Baseline (κ)	Baseline (MAE)	Full (κ)	Full (MAE)
Final Grade	.4226	161.57	.4332	160.62
Formal Language	.3690	35.00	.3834	34.92
Understanding the task	.3507	37.82	.3624	37.69
Organization of Information	.3427	37.91	.3557	37.73
Knowing argumentation	.3318	40.79	.3479	40.58
Solution proposal	.2835	43.13	.2987	42.83

When the value of κ is closer to 1, the higher the agreement between evaluators, and when the value of κ is closer to 0, the lower the agreement between evaluators. MAE is a more intuitive metric, and due to this advantage, we employed it in our tests as well.

3.3.1 An analysis of handcrafted features

The analysis of handcrafted features is concerned with RQ1. To perform such an analysis, we applied Linear Regression (LR) to predict the final grade of essays, and each of the other five aspects. Also, a simple oversampling strategy is applied since grade distribution is unbalanced (Figures 3.1 and 3.2). In the first step of oversampling strategy, it searches by the class G_{max} that holds the largest number of instances, where each class corresponds to a score (3.2). Then, the strategy randomly selects instances from every class $G \neq G_{max}$ and replicates such instances into training datasets, until the size of every class $G \neq G_{max}$ is equal to the size of G_{max}.

Tables 3.5 and 3.6 describes the κ's results for both our datasets using our baseline configuration or the full configuration (domain + baseline features) and simple linear regression. We executed cross-validation five times and compute the average of the κ of all experiments, for each aspect and final grade, to evaluate oversampling performance. Again, we assessed the statistical significance of our measurements by comparing each pair of models using Welch's t-test with a p$-$value ≤ 0.01.

Considering the lack of tools for processing the Portuguese language, and the limited performance of the few existing tools, the multi-aspect classification performed satisfactorily. However, some aspects performed poorly, probably due to the subjectivity intrinsic to these aspects, and objective variables perhaps can not capture all the subjectivity.

Table 3.6. Results for each grade aspect for Brasil Escola Dataset

Grade Type	Baseline (κ)	Baseline (MAE)	Full (κ)	Full(MAE)
Final Grade	0.2524	146.67	0.2667	145.91
Formal Language	0.2848	27.3831	0.2889	27.31
Understanding the task	0.2972	31.39	0.3078	31.24
Organization of Information	0.2192	39.74	0.2369	39.44
Knowing argumentation	0.1923	37.54	0.2037	37.36
Solution proposal	0.2163	42.15	0.2324	41.92

3.3.1.1 Individual Feature Analysis

A full investigation of the influence of features on a model comprises tests of all possible feature combinations. However, the number of tests would take a long time, and some of the tests would lead to irrelevant results to study. Thus, we tested possible features sets using a technique called dynamic programming. This technique builds the optimal final solution based on smaller suboptimal solutions, i.e., it divides the problem into small problems and tries to find optimal solutions for these small problems to produce the final solution. Applying this concept to our context, we aimed to select a set of features that produces the best value of κ. Thus, considering the whole set of features as S, at some time t, we are testing a subset of features S_f^t, which contains a feature f. If the κ_f^t value for the subset S_f^t is lower than the previous solution, then we discard f and go to the next round test other features. Otherwise, f is included in our solution, and we go to the next round as well. This process goes on until all features are tested.

The learning strategy for these tests was the Linear Regression (LR) algorithm, and we generated a combination of features to leverage the complexity of our model. In the combination of features, for two features f_1 and f_2, we build a new feature $f_2^1 = f_1/f_2$. In this scheme, the experiments tested 240 features.

Now, consider the following set $\phi_f = \{\kappa_f^t | \text{t is t-h round}\}$. We computed the average of ϕ_f for each $f \in S$ and then sorted in ascending order. Afterward, we can easily find the features that most influence the model, and with the largest κ. First, we compared the datasets in each aspect, and then we analyzed the results by aspect in each model.

Comparison between Datasets. For each aspect, we investigate if there are differences or similarities between UOL and Brasil Escola datasets. We list only three most relevant and three least relevant features to perform a clean analysis of our results.

Table 3.7 shows the most and the least relevant features for the formal language aspect. The most influential features in the Brasil Escola dataset are related to this

Table 3.7. Individual analysis of features of Formal Language (FL) Score: UOL and Brasil Escola datasets

Category	Dataset	Feature	κ
Most Relevant	UOL	NumberDiffWords/RacistTerms	.4158
		RacistTerms	.4158
		SpellingCheckNorm/SubordinateClause	.4158
	Brasil Escola	GrammarErrorsNorm	.3205
		GrammarErrorsNorm/AnaphorPronouns	.3204
		SentenceLong/AnaphorPronouns	.3204
Least Relevant	UOL	GrammarErrors/SentenceLong	.3923
		Enclise/DemonstrativeNorm	.3938
		SimilarityPrompt/EncliseNorm	.3966
	Brasil Escola	StyleErrors/NumberDiffWords	.2628
		NumberTokens/RacistTerms	.3016
		SpellingCheckNorm/AnaphorPronouns	.3016

Table 3.8. Individual analysis of features of Understanding the Task (UT) Score: UOL and Brasil Escola datasets

Category	Dataset	Feature	κ
Most Relevant	UOL	Enclise/DiscourseMarkers	.4114
		Enclise/SimilarityPrompt	.4114
		SubordinateClause/AnaphorPronouns	.4114
	Brasil Escola	GrammarErrors/NumberDiffWords	.3282
		SentenceLong/RacistTerms	.3282
		FleshScore	.3282
Least Relevant	UOL	NumberDiffWords	.3645
		NumberTokens/SentenceLong	.3766
		NumberTokens/NumberDiffWords	.3820
	Brasil Escola	NumberDiffWords	.2830
		GrammarErrors	.3200
		Enclise/SpellingCheck	.3232

aspect, especially GrammarErrorsNorm. However, for the UOL model, the most relevant features for this aspect are missing, except for SpellingCheckNorm. Perhaps, the formal language evaluation in the Brasil Escola dataset is more decisive than in the UOL dataset. With respect to the least relevant features, we observe a few similarities between the two models, and maybe the combination of such features is unrelated to formal language.

For the aspect Understanding the Task, Table 3.8 shows that the UOL model

Table 3.9. Individual analysis of features of Organization of Information (OI) Score: UOL and Brasil Escola datasets

Category	Dataset	Feature	κ
Most Relevant	UOL	SimilarityPrompt	.3953
		StyleErrors	.3952
		Enclise/DiscourseMarkersNorm	.3952
	Brasil Escola	Demonstrative/SentenceLong	.2658
		FleshScore	.2658
		NumberTokens/SentenceLong	.2658
Least Relevant	UOL	NumberDiffWords	.3563
		StyleErrors/SimilarityPrompt	.3685
		Demonstrative/FleshScore	.3891
	Brasil Escola	GrammarErrors/SentenceLong	.2124
		NumberDiffWords	.2411
		GrammarErrors/SimilarityPrompt	.2552

employs more refined features to predict the understanding score. We highlight the relevance of *ênclise*'s features that are in the first and the second places. Regarding the least important features, NumberDiffWords seems to play an insignificant role in this aspect. Perhaps, a variety of vocabulary is not important to express understanding of the task.

Regarding the aspect Organization of Information, none of the relevant features match. However, the two models present coherent features. For instance, SimilarityPrompt in the UOL model reveals that the adequate selection of information influences this score. In addition to that, the feature FleshScore in the Brasil Escola model unveils that high schooling is necessary to organize a text properly. Considering the least relevant features, we highlight the presence of NumberDiffWords in both models. Perhaps, to employ an appropriate vocabulary is enough for this aspect, and the student does not have to diversify her vocabulary.

The feature analysis for the aspect Knowing Argumentation is depicted in Table 3.10. Concerning the Brasil Escola model, the best results are achieved employing grammar features, which is a surprising result, since to know argument requires most sophisticated features. A similar result is presented in the UOL model; nonetheless, some non-grammar features are present, like SimilarityPrompt. Perhaps, in this dataset, the vocabulary should be related to the topic of the prompt. With respect to the least relevant features, there are many differences. While for UOL, FleshScore is a distinctive feature that is unimportant, the DemonstrativeNorm feature is irrelevant for the Brasil Escola model. Surprisingly, discourse markers are not present in none of

Table 3.10. Individual analysis of features of Knowing Argumentation (KA) Score: UOL and Brasil Escola datasets

Category	Dataset	Feature	κ
Most Relevant	UOL	SpellingCheck/SimilarityPrompt	.3804
		SpellingCheck	.3803
		DemonstrativeNorm/LexicalComplexity	.3803
	Brasil Escola	GrammarErrorsNorm	.2277
		NumberDiffWords/StyleErrorsNorm	.2277
		Demonstrative/EncliseNorm	.2277
Least Relevant	UOL	StyleErrors/SimilarityPrompt	.3468
		FleshScore	.3597
		FleshScore/NumberDiffWords	.3650
	Brasil Escola	DemonstrativeNorm	.2136
		GrammarErrors/NumberDiffWords	.2175
		StyleErrors/DemonstrativeNorm	.2232

Table 3.11. Individual analysis of features of Solution Proposal Score (SP): UOL and Brasil Escola datasets

Category	Dataset	Feature	κ
Most Relevant	UOL	FleshScore/SimilarityPrompt	.3416
		Enclise/SimilarityPrompt	.3415
		StyleErrorsNorm	.3415
	Brasil Escola	GrammarErrors/StyleErrorsNorm	.2575
		SpellingCheck/RacistTerms	.2575
		SimilaritySentences	.2575
Least Relevant	UOL	StyleErrors/SimilarityPrompt	.3016
		NumberDiffWords	.3017
		FleshScore/NumberDiffWords	.3102
	Brasil Escola	GrammarErrors/SimilarityPrompt	.2309
		StyleErrors	.2340
		EncliseNorm/SpellingCheckNorm	.2451

the models.

Table 3.11 details the results for the Solution Proposal aspect. A feature developed to this aspect, ConclusionMarkers, is missing from both models. However, in the UOL model, the FleshScore feature is listed as relevant, which is intuitive since it captures the school degree of the student. Regarding the remaining relevant features, no clear explanation is possible to develop. Perhaps that is why the results are low compared to the other aspects. About the least relevant features, we observe that

Table 3.12. Individual analysis of features of Final Grade(FG): UOL and Brasil Escola datasets

Category	Dataset	Feature	κ
Most Relevant	UOL	SentenceLong/ENcliseNorm	.4592
		FleshScore/EncliseNorm	.4592
		EncliseNorm/DiscourseMarkersNorm	.4592
	Brasil Escola	StyleErrorsNorm/RacistTerms	.2925
		NumberTokens/StyleErrorsNorm	.2925
		DiscourseMarkers/LexicalComplexity	.2925
Least Relevant	UOL	SentenceLong/GrammarErrorsNorm	.4443
		Enclise/FleshScore	.4531
		Demonstrative/FleshScore	.4532
	Brasil Escola	FleshScore/SimilarityPrompt	.2300
		GrammarErrors/SimilarityPrompt	.2699
		NumberTokens/FleshScore	.2713

Table 3.13. κ results for the ablation tests

	UOL (abl./full)		Brasil Escola (abl./full)	
Aspect	κ	#ftrs	κ	#ftrs
Final Grade (FG)	.4615/.4332	33/30	.2950/.2667	49/30
Formal Language (FL)	.4185/.3834	32/30	.3237/.2889	49/30
Understanding the task (UT)	.4139/.3624	38/30	.3296/.3078	41/30
Organization of Information (OI)	.3973/.3557	24/30	.2675/.2369	46/30
Knowing argumentation (KA)	.3828/.3479	24/30	.2296/.2037	44/30
Solution proposal (SP)	.3439/.2987	27/30	.2595/.2324	45/30

StyleErrors/SimilarityPrompt and GrammarErrors/SimilarityPrompt are related. Although we can think that an essay that is similar to the prompt and presents fewer grammar or style errors is better, this idea is uncorrelated with this aspect. Probably because it is expected that the student presents ideas clearly, which can only be captured by more complex features.

Comparison between Aspects.

The best models for each aspect are reported in Table 3.13. Brasil Escola dataset requires many more features to build a robust dataset, while the UOL dataset in three aspects even involves fewer features than the full feature set.

Most of the combinations of the features are in only one aspect. Nonetheless, some of them are in at least five aspects or even in the final score. These features seem crucial to their model; therefore, it is important to look at them. The following list details the frequent features in the aspects.

- **UOL**
    1. DiscourseMarkersNorm/SpellingCheckNorm
    2. DiscourseMarkers/GrammarErrorsNorm
    3. SimilarityPrompt
    4. WordLength
    5. IndeterminationInstrucments
- **Brasil Escola**
    1. SimilarityPrompt/NumberDiffWords
    2. SentenceLong
    3. RacistTerms
    4. StyleErrors/NumberDiffWords
    5. WordLength
    6. SpellingCheck/EncliseNorm
    7. IndeterminationInstrucments
    8. LexicalComplexity
    9. DemonstrativeNorm/DiscourseMarkersNorm
    10. GrammarErrorsNorm
    11. StyleErrors/DiscourseMarkersNorm

In this list, we see that the feature DiscourseMarkers plays an essential role in all the aspects and final grade. If the writer is unable to go from an idea to another, then the evaluator is incapable of performing a valuable assessment. Also, the context of employed vocabulary, represented by the feature SimilarityPrompt, is relevant for the evaluation. Maybe this is why most of the models proposed by other authors are built for each prompt instead of a prompt-unaware model.

Besides the comparison between aspects in each dataset, we also compared the aspects of datasets. For this comparison, we inspect the features that are in both models in the same aspect. Hence, it is possible to verify the relevant features to an aspect in general. Next, we detail which features are in both models categorized by aspect.

Final Grade. Errors related to the formal language, like StyleErrors, seems important for the Final Grade. Also, vocabulary (NumberDiffWords) and the way the text is organized (Conjunctions) are relevant to the final score. A good vocabulary, the style of the text, and the organization of text are essential for any reader understanding a text. Whether these elements are missing from the text, perhaps it is hard to evaluate more sophisticated features.

1. NumberDiffWords
2. WordLength
3. IndeterminationInstrucments
4. Conjunctions
5. StyleErrors/StyleErrorsNorm
6. SentenceLong/NumberDiffWords

Formal Language. As expected, some grammar features are in the Formal Language list (StyleErrorsNorm and StyleErrors/SimilarityPrompt). Other features that capture the use of punctuation (SentenceLong/EncliseNorm and SentenceLong/AnaphorPronouns) are also in the list. Curiously, RacistTerm is in both models of formal language. One explanation is that some of the words or expressions in racist terms list can be unrelated to the racial offenses in essays since the racial terms were collected from social media texts.

1. StyleErrorsNorm
2. NumberDiffWords
3. WordLength
4. Articles
5. SimilarityPrompt
6. RacistTerms
7. StyleErrors/SimilarityPrompt
8. SentenceLong/EncliseNorm
9. FleshScore/SimilarityPrompt

10. SentenceLong/AnaphorPronouns

Understanding the Task. As expected, if the writer employs a vocabulary similar to the prompt, and varies the words (SimilarityPrompt/NumberDiffWords) are good indicators that the writer understood the prompt. The vocabulary is another influential feature as well (StyleErrors/NumberDiffWords), which corroborates the relevance of vocabulary for this aspect. Again, grammar features (GrammarErrorsNorm/SubordinateClause and StyleErrors/NumberDiffWords) are in another aspect. We believe that such grammar features are tools for the student to be clear in her writing, and that is why some of these features are in most aspects.

1. WordLength
2. IndeterminationInstrucments
3. GrammarErrorsNorm/SubordinateClause
4. DemonstrativeNorm/DiscourseMarkersNorm
5. SimilarityPrompt/NumberDiffWords
6. StyleErrors/NumberDiffWords
7. Demonstrative/DiscourseMarkersNorm

Organization of Information. We could suppose that features like Conjunctions and DiscourseMarkers are a consensus between evaluators. However, they miss from the following list. On the other hand, the similarity with a prompt is a regular feature (SimilarityPrompt, SimilarityPrompt/NumberDiffWords, SimilarityPrompt/NumberDiffWords). Probably, this feature is significant for this aspect because it considers the selection of relevant information by the student.

1. WordLength
2. IndeterminationInstrucments
3. SimilarityPrompt
4. SentenceLong/GrammarErrorsNorm
5. NumberTokens/SimilarityPrompt
6. SimilarityPrompt/NumberDiffWords

7. StyleErrors/NumberDiffWords

8. GrammarErrorsNorm/StyleErrorsNorm

9. FleshScore/SubordinateClause

Knowing Argumentation. This aspect presents the smallest set of features in common. Possibly, the subjectivity of this aspect results in a discrepancy between evaluations. Nevertheless, we observe that even with a small set of features, the vocabulary (NumberDiffWords and NumberDiffWords/StyleErrorsNorm) is a frequent attribute in the following list.

1. NumberDiffWords

2. WordLength

3. IndeterminationInstrucments

4. NumberDiffWords/StyleErrorsNorm

Solution Proposal. This aspect requires that the student respect human rights, then although RacistTerms is an inaccurate list for our domain, it seems to be somehow relevant (StyleErrorsNorm/RacistTerms). StyleErrorsNorm is a frequent feature in this aspect as well (StyleErrorsNorm, StyleErrors/NumberDiffWords, StyleErrorsNorm/RacistTerms). Maybe, how the student presents her solution can harm or leverage the score.

1. StyleErrorsNorm

2. IndeterminationInstrucments

3. SimilarityPrompt

4. SimilarityPrompt/NumberDiffWords

5. StyleErrors/NumberDiffWords

6. Demonstrative/DiscourseMarkersNorm

7. Demonstrative/SentenceLong

8. StyleErrorsNorm/RacistTerms

We also highlight the IndeterminationInstruments feature, which is in almost all aspects. The components of this feature are the use of passive voice, the use of the first singular person, and some kind of *ênclise*. All these elements are related to the formal use of the language. One possible reason for this feature stands out is the limitations of the tools of other formal language features.

For more details, we refer the reader to Appendix E with the complete lists of features for all aspects and datasets.

Although individual analysis of features is fascinating, LR is not a robust method and does not show the complex interaction between features. Then, in the next section, we present a more sophisticated analysis of our strategy.

3.3.1.2 SHAP analysis

The SHapley Additive exPlanation (SHAP) is an approach to explain the output model of some machine learning algorithms, like XGboost [Lundberg and Lee, 2017]. The advantage of the SHAP approach is that it provides an understanding of more complex models than, for example, LR. In this technique, Shapley values yield the analysis. According to game theory, these values are the amount of responsibility of each agent in the success of a collaborative game.

Using these values, the SHAP library[9] allowed us to plot for each aspect the feature importance for XGboost learning approach. In our experiments, we performed 5-fold cross-validation for each aspect in our datasets, and then we concatenated the SHAP values for each fold. Finally, the graph of the feature importance is plotted.

Final Grade. For the final grade, we can analyze Figure 3.7. The graphs depict the feature importance for our datasets using the full feature set. On the right side of each graph, there is a scale that points out that the higher the feature value, the more similar to the red color the data point is. The scale at the bottom of the graph represents the impact of the feature in the model. The impact can be positive or negative. For instance, consider the number of errors per token in the first position in the graph, we can see that the red data points lead to a lower prediction than the blue data points, which leads to higher predictions.

It is possible to observe that the importance of the features is different for each evaluator. In education, this is a well-known phenomenon that some evaluators favor some features rather than others [Freedman, 1979]. Napoles and Callison-Burch [2015] analyzed the weights of the linear regression model employed to score essays. However, in our study, we use a more sophisticated method. According to the graphs, the

[9] `https://github.com/slundberg/shap`

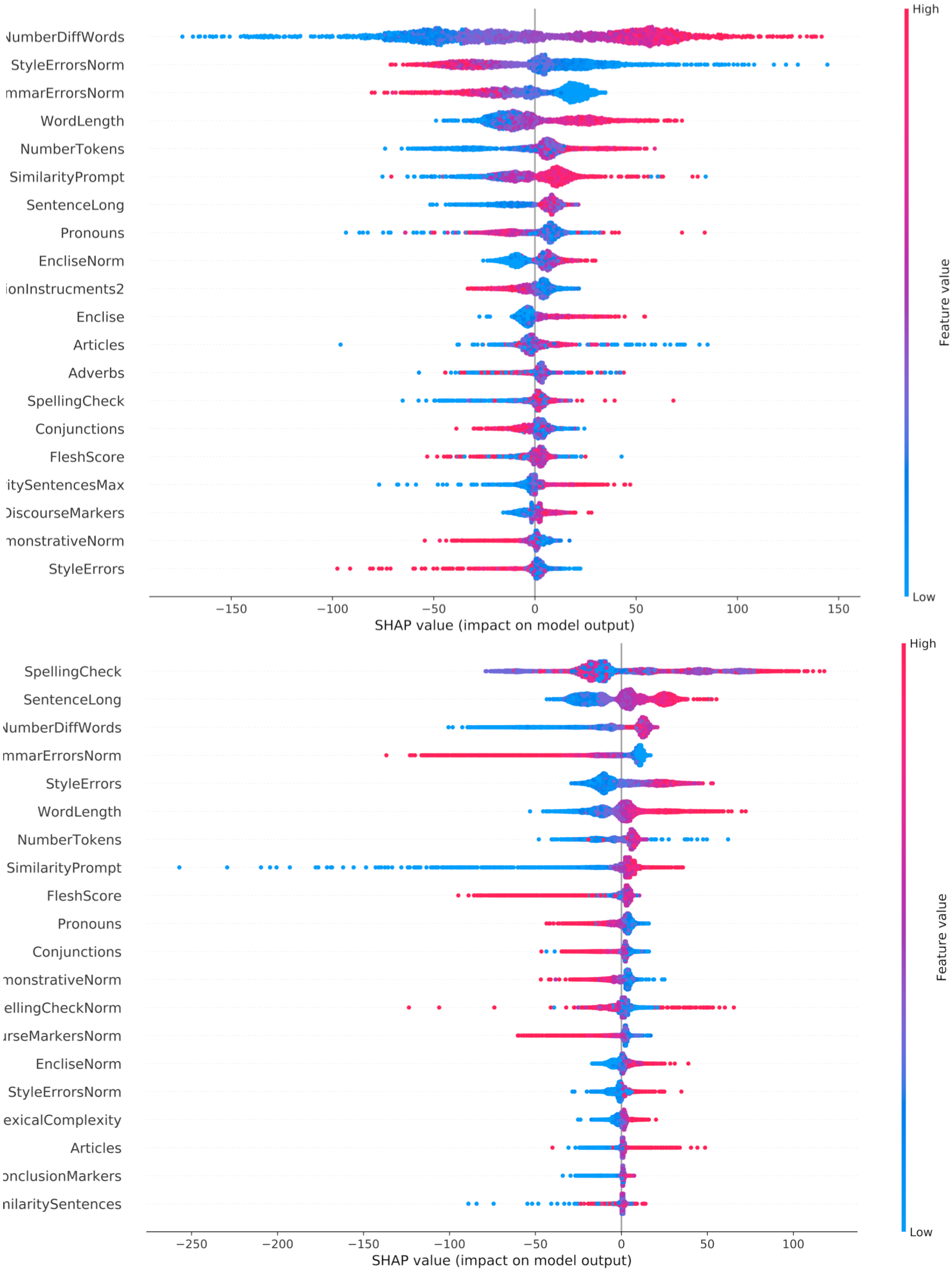


Figure 3.7. SHAP graphs showing the feature importance for the Final Grade of UOL Dataset (up) and Brasil Escola dataset (bottom)

evaluators agree in almost half of their top ten features, namely: NumberDiffWords, WordLength, SentenceLong, SimilarityPrompt, GrammarErrors, and StyleErrors. It is reasonable to think that although evaluators assess features differently, there is a consensus about the relevant ones. However, some of the features are disagreement between evaluators. While ConclusionMarkers seems to have some importance in the Brasil Escola, it does not even appear in the UOL graph. Also, we can point to the Adverbs that shows in the UOL and is missing in the Brasil Escola. Not to mention the SpellingCheck feature that is almost in the last position in the UOL while it is in the first position in the Brasil Escola. Lastly, it is possible to perform numerous analyses in these graphs, and they corroborate several studies of the educational area.

In addition to that, we would like to highlight that in the full feature set system, the Anaphor's features represented leverage in the results. The baseline graphs can be checked in Appendix D.

Formal Grade. Figure 3.8 displays the feature importance for the Formal Grade aspect for UOL and Brasil Escola dataset. We can notice that some features in both graphs match with features in the Final Grade graphs. For instance, in the Formal Grade, StyleErrors are more relevant for the UOL model than for the Brasil Escola model. SpellingCheck still is an important feature for the Brasil Escola model; however, its influence is less than in the final score. Also, there are a few differences between the formal grade graphs. These differences are probably due to the objectivity of the Formal Grade skill.

Understanding the Task. The feature importance of Understanding the Task is described in the graphs of Figure 3.9. NumberDiffWords is in a higher position than in the other aspects. Perhaps the vocabulary is more relevant for this skill than Formal Language because of showing how you understand the task. Also, you should use a diversified vocabulary and, at the same type words that are similar to the vocabulary domain. Also, that is why SimilarityPrompt is relevant to this skill in both datasets. In addition to these similarities, there are some differences besides the ranking of the features. IndeterminationInstrucments and Demonstrative are in the Brasil Escola model while they are missing in the UOL model. Indetermination instruments allow more diversity in sentence construction, and demonstrative pronouns can have an anaphoric role. Both features are related to a more formal language, which can be an aspect that the evaluator of the Brasil Escola grants more importance when the student is presenting her understanding of the task. On the other hand, the feature SimilaritySentences is only listed in the UOL model, which can mean that this evaluator weights more the cohesion of ideas in Understanding Task skill.

Organization of Information Grade.

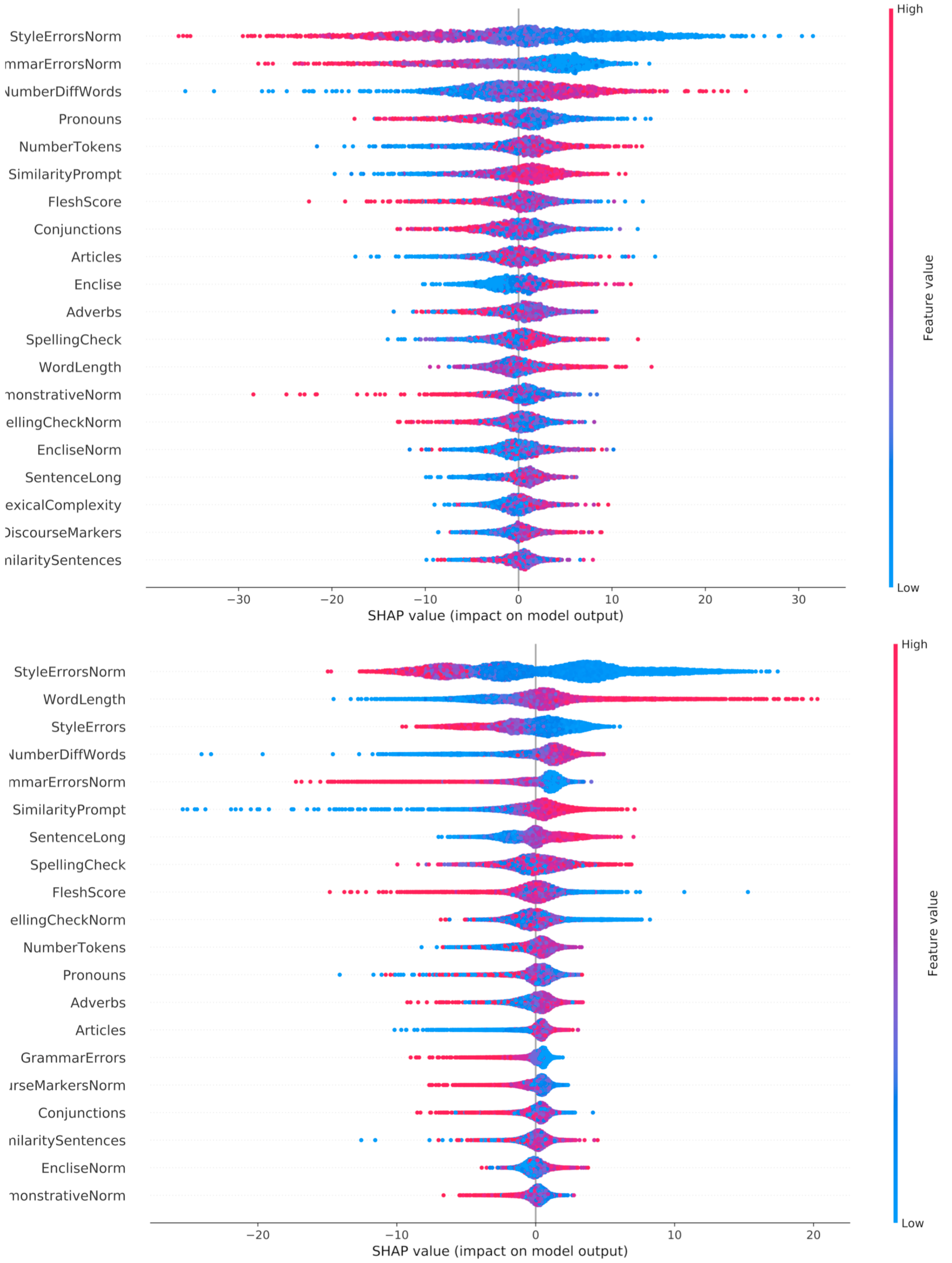


Figure 3.8. SHAP graphs showing the feature importance for the Formal Grade of UOL Dataset (up) and Brasil Escola dataset (bottom)

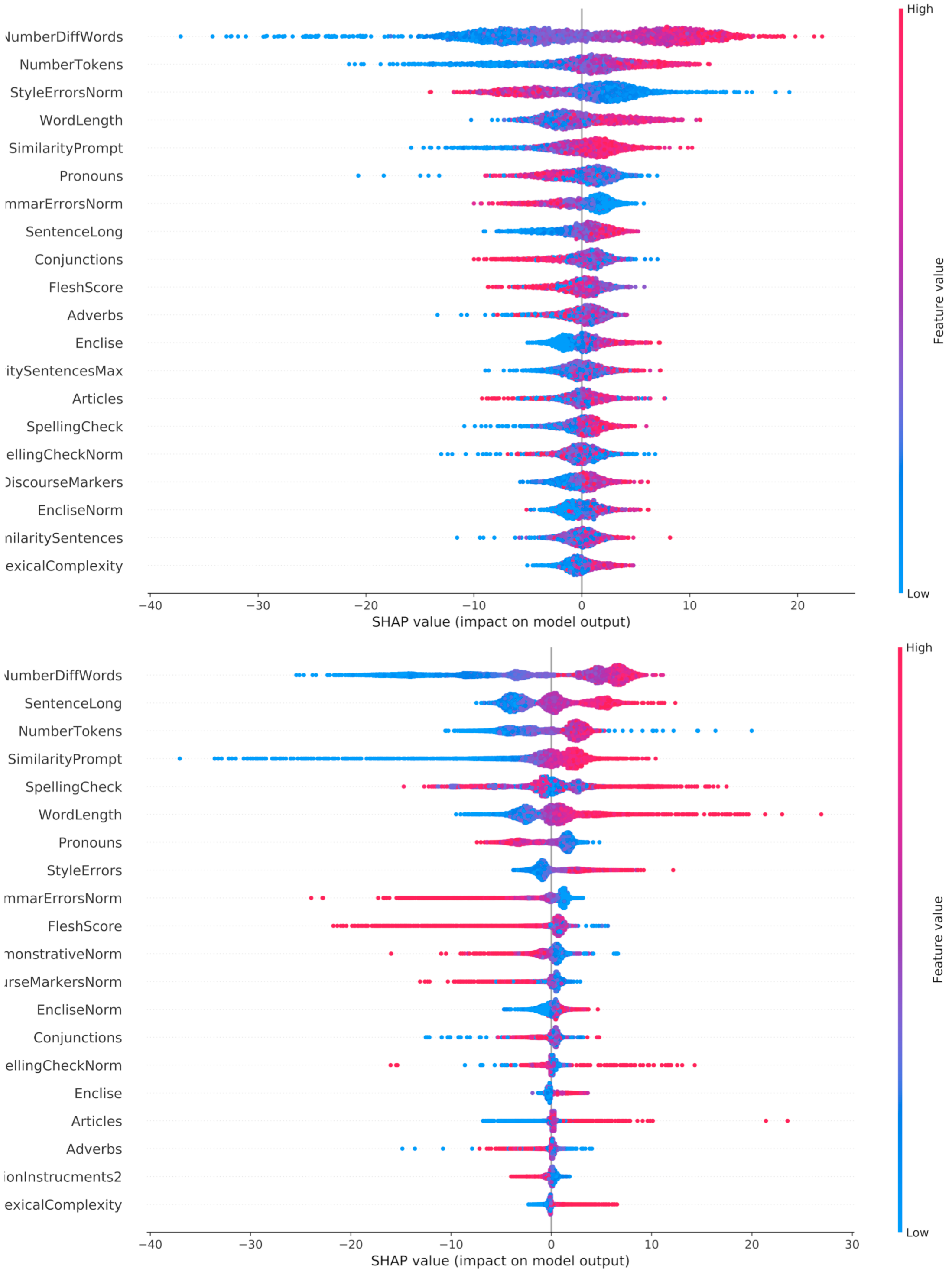


Figure 3.9. SHAP graphs showing the feature importance for the Understanding the Task grade of UOL Dataset (up) and Brasil Escola dataset (bottom)

Figure 3.10 presents the feature importance ranks for the Organization of information aspect in our datasets. When analyzing the graphs of the Organization of Information, we expect that some features would present more relevance, like the discourse markers that aim to link the text coherently. And indeed, in the UOL model, we can notice that the DiscourseMakers achieved a higher position than in the final score graph. However, when we look at the Brasil Escola model, this change is not present, and if we compare the final score graph with the organization of the information graph, we notice that there are few differences between the features lists. Perhaps the lack of change between scores is that the evaluator can be assessing the skill of the Organization of Information in a generic way. Back to the UOL model, we note another divergence compared to the Final Score graph, the absence of the LexicalComplexity feature. Since this feature is related to the formal language aspect, the UOL evaluator may expect that the student organizes the information using more formal writing. Finally, we could presume that anaphoric elements would be relevant to this skill. Nevertheless, there is no change regarding these elements compared to the Final Grade.

Knowing Argumentation. For Knowing Argumentation aspect, Figure 3.11 presents the feature importance graphs. According to Tables 3.5 and 3.6, Knowing Argumentation is one of the skills with the lowest performance. Perhaps this is due to the few features of our model related to this skill. This skill is especially hard to develop features because it requires labeled data (Madnani et al. [2012], Stab and Gurevych [2014], Stab and Gurevych [2016]). Because of the lack of features to describe this skill, there are few differences between the Final Score and the Knowing Argumentation score. As future work, we suggest to label elements as major claims, claims, and premises in the datasets that we presented.

Solution Proposal.

The Solution Proposal aspect, Figure 3.12, like Knowing Argumentation skill, presented a poor performance. Nonetheless, its main feature, ConclusionMarkers, seems to influence the result in the UOL model, as we can observe in Figure 3.12. While in the Final Score it is missing in the graph, in the solution proposal graph it is present. The discourse markers features, DiscourseMakers, and DiscourseMarkersNorm are also in the UOL model. The presence of these features shows the importance of these features for this skill. It is not possible to see such differences in the Brasil Escola model, though, and we believe that is because the Brasil Escola evaluator is doing a generical evaluation of this skill as well. In addition to that, we can highlight that our model lacks features related to this skill. We suggest as future work, make comparisons with high grades essays, and maybe the distance between an essay and high-grade essay to

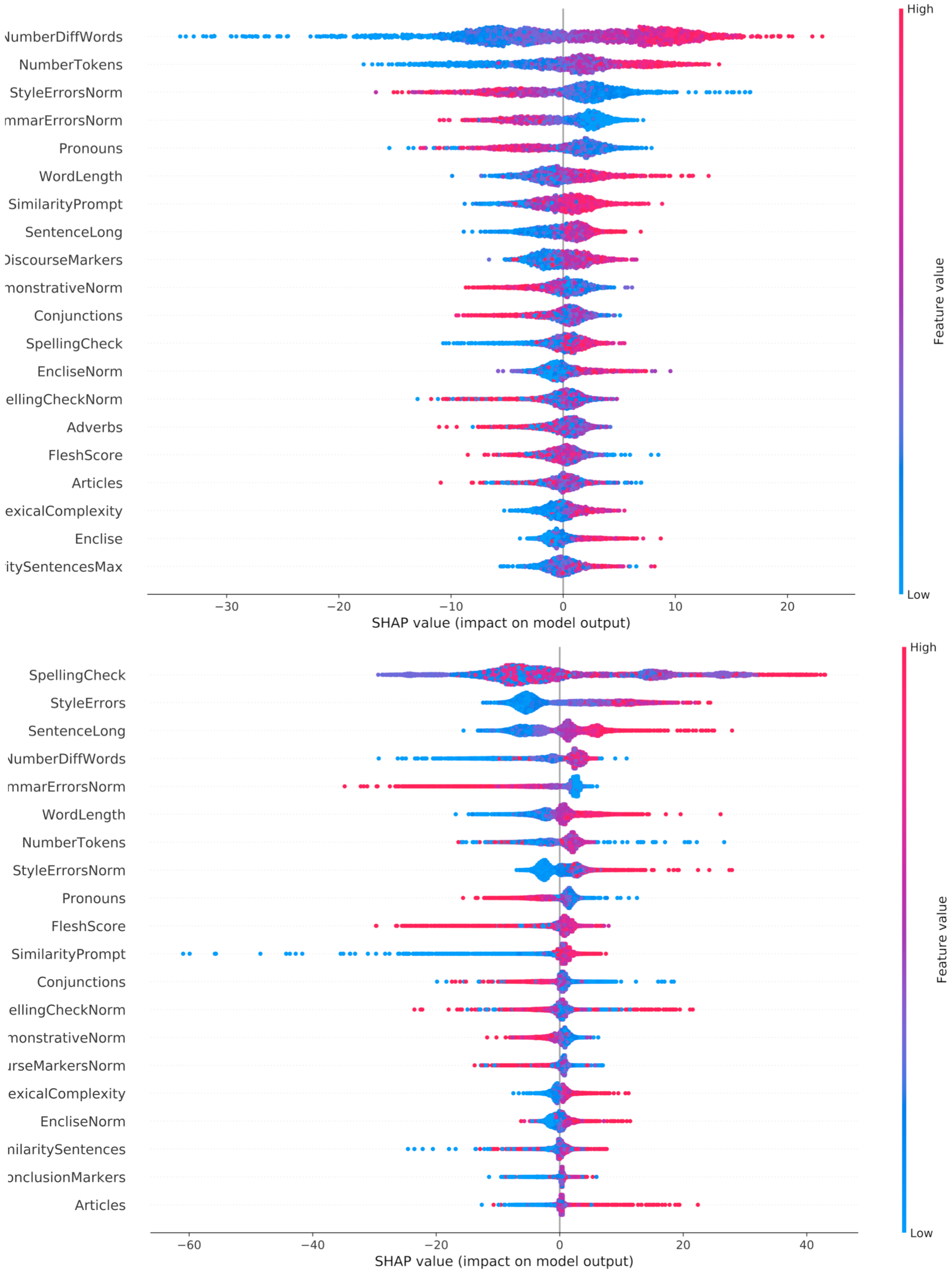


Figure 3.10. SHAP graphs showing the feature importance for the Organization of Information Grade of UOL Dataset (up) and Brasil Escola dataset (bottom)

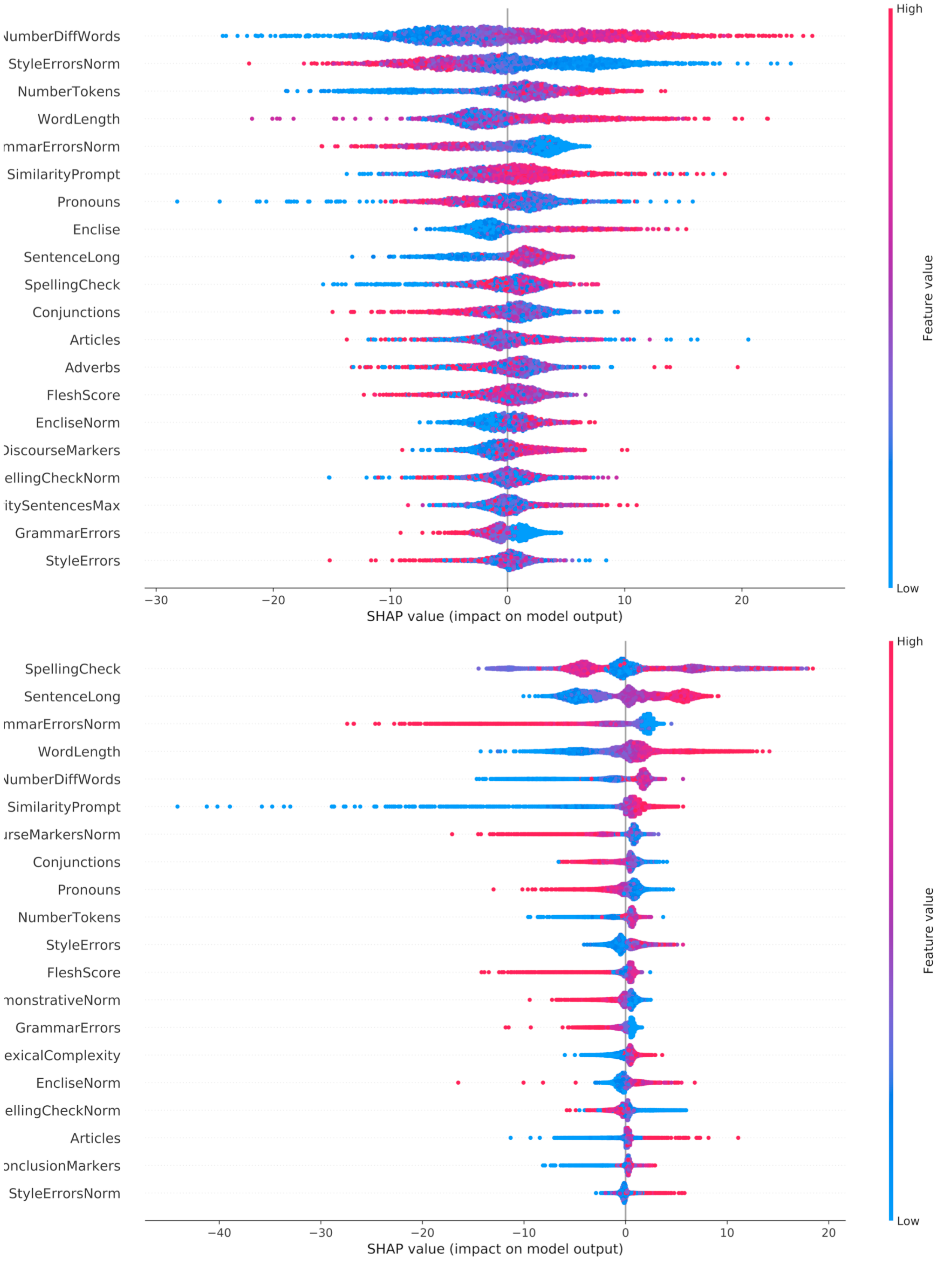


Figure 3.11. SHAP graphs showing the feature importance for the Knowing Argumentation grade of UOL Dataset (up) and Brasil Escola dataset (bottom)

capture the solution proposal differences.

3.3.2 Comparison of machine learning techniques

We learn AES models using Random Forests (RF), Linear Regression (LR), Gradient Boosting (GB), and Multi-Layer Perceptron (MLP). All models are based on the same set of features, previously described in Section 3.3.1, and all models are trained in the regression model. The metric used to evaluate the effectiveness of the different models is the quadratic weighted kappa (κ). 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 the training set, and the remaining fold is used as the 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 hyperparameters, and the test set is used to estimate κ numbers for the different 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 assessed the statistical significance of our measurements by comparing each pair of models using Welch's t-test with a $\mathrm{p-value} \leq 0.01$.

Score distribution: This experiment is concerned with RQ2. Figure 3.13 shows how scores are distributed over the essays in our corpora. Although the distribution differs for each AES model, scores in the UOL dataset are centered around 400, 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. The classifiers trained in the Brasil Escola dataset followed the trend of the human evaluator to assign higher grades than the human evaluator of the UOL dataset, then the scores for the Brasil Escola are centered around 600. Again, LR behaved similarly and assigned lower grades than other classifiers, and GB and MLP also assigned better-distributed scores.

Figure 3.14 shows how aligned with human raters are the different AES models. For most of the essays, AES models are well aligned with human raters, showing misalignment that varies from -200 to $+200$. When there is a difference higher than +/-200, the tendency is that this difference is negative instead of positive. Therefore there is a negative bias in the classifiers.

Comparison between techniques: This experiment aims to answer RQ3. Tables 3.14 and 3.15 describe the results of all techniques to UOL and Brasil Escola datasets.

As it is possible to observe, the best performances for the UOL dataset alternated between GB and LR, while in the Brasil Escola dataset, the best results are from the

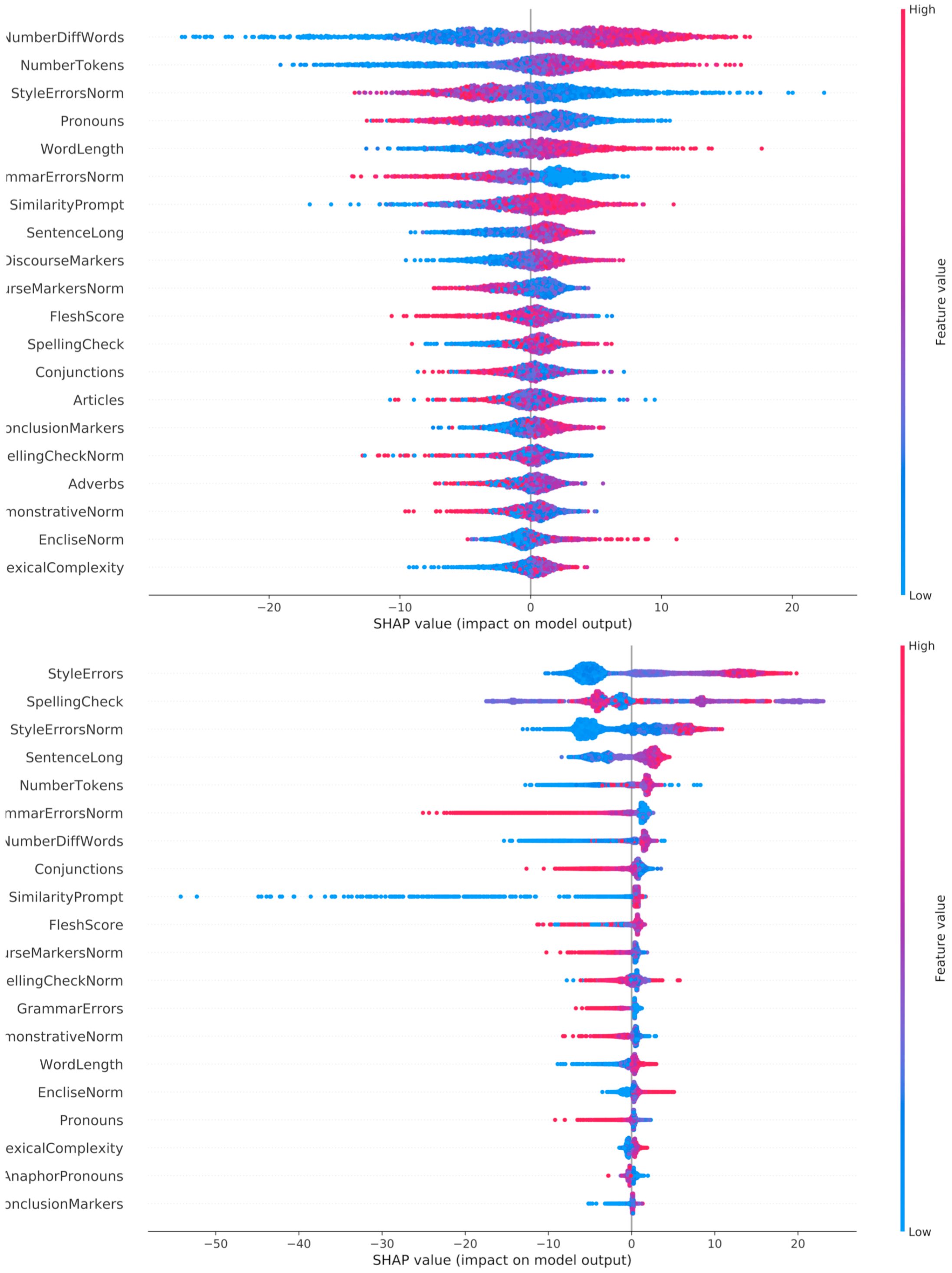


Figure 3.12. SHAP graphs showing the feature importance for the Solution Proposal grade of UOL Dataset (up) and Brasil Escola dataset (bottom)

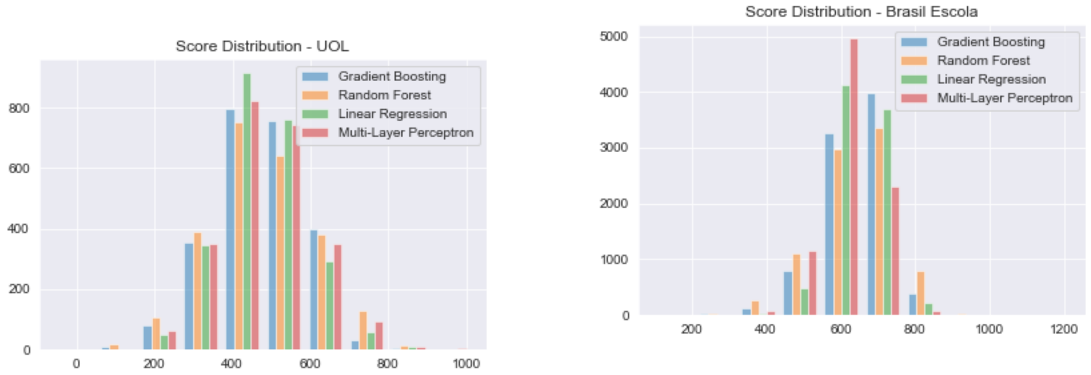


Figure 3.13. Distribution of the scores given by different AES models (UOL/Brasil Escola).

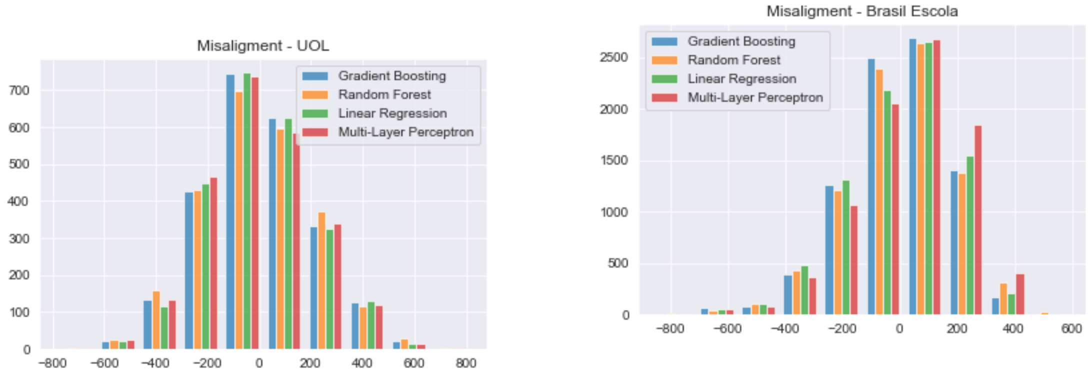


Figure 3.14. Distribution of misalignment for the different AES models (UOL/Brasil Escola).

GB. Another remark is that we developed a few features for the skills Organization of Information and Solution Proposal. Nonetheless, both presented a higher performance in the Brasil dataset than in the UOL dataset. These results confirm the explanations of the Section 3.3.1.2 that reveal minor changes between the final score and these skills. Probably, the proposed features were enough to build robust models for the Organization of Information and Solution Proposal, even without enough features to describe them. However, it is best to develop specific features to these aspects since a transparent model is desirable for the AES task.

In addition to that, structured techniques performed better than the non-structure technique, MLP. Perhaps this is because the proposed features are simple, and a more robust technique requires more complex representations, like word embeddings.

Algorithm	Final Grade	Understanding the task	Organization of information	Knowing argumentation	Solution proposal	Formal Language
GB	.4040	**.3566**	.3435	**.3267**	.2802	.3652
RF	.3604	.2976	.3035	.2983	.2398	.3129
LR	**.4204**	.3537	.3423	**.3267**	**.2885**	**.3721**
MLP	.3923	.3298	**.3498**	.3066	.2625	.3359

Table 3.14. κ numbers for different models UOL

Algorithm	Final Grade	Understanding the task	Organization of information	Knowing argumentation	Solution proposal	Formal Language
GB	**.3819**	**.3553**	**.3896**	**.2823**	**.4128**	**.3326**
RF	.3246	.2826	.3609	.2329	.3656	.2886
LR	.2501	.2888	.2148	.1815	.2140	.2776
MLP	.2614	.2929	.2822	.1896	.3095	.2811

Table 3.15. κ numbers for different models Brasil Escola

Also, we can notice that Formal Language is the aspect that should perform better due to the features proposed. However, for the Brasil Escola dataset, the aspect with the highest performance is the Solution Proposal. On the other hand, the UOL dataset performed as we expected, and Formal Language presented the highest κ among the aspects.

3.3.3 Discussion

At the beginning of this chapter, we presented three research questions. The first question is about which features are essential for essay scoring when we employ Linear Regression (LR) as a learning technique. To answer this question, we performed an ablation experiment, in which each round one feature is ablated from the full set of features. For each feature $f \in F$, we took a set of models M_f that f is included, and we computed the average of κ's. We called this average of κ_f, for some given feature f. The following list $L_\kappa = \{\kappa_f, \forall f \in F\}$ were sorted in descending order to evaluate the influence of the features in each aspect. We demonstrated using ablation experiments and the subsequent analysis that Baseline features presented a higher impact than the Domain features. This result is confirmed by the preliminary results of Tables 3.5 and 3.6, which both showed robust results for the Baseline features. These outcomes are especially evident for the Formal Language aspect since several Baseline features are related to it.

Regarding the Organization of Information aspect, we also demonstrated that the most influential features are related to this aspect, which are DiscourseMakers, DiscourseMarkersNorm, and SentenceLong. These features are Baseline features, and the Domain features are missing from the table of Organization of Information, Table 3.9. However, we performed a more detailed analysis, and the Domain features also affect the Organization of Information. For instance, SimilaritySentences are relevant to the UOL dataset, and Conjunctions and Pronouns are relevant to the Brasil Escola model. The same reasoning we applied to Knowing Argumentation aspect since DiscourseMarkers and IndeterminationInstruments are important for both datasets. Finally, we developed only two specific features for the Solution Proposal aspect. Again, they are missing from the Solution Proposal table (Table 3.11), but a more detailed analysis revealed to us that ConclusionMarkers are influential to the UOL dataset, and RacistTerms are influential to Brasil Escola dataset. Using this process, we answered our first research question. Moreover, we showed that specific features for each aspect positively influence the prediction, which suggests that a more sophisticated feature engineer leverage the results for each aspect. Thus, our first question was answered in Section 3.3.1 by numerous experiments and analysis.

Regarding the second question, Figure 3.13 shows that the algorithms, in general, are biased towards the scoring in the middle grades. This bias is probably due to the imbalanced dataset, which is composed mostly of average scores. It is a well-known phenomenon in education that human evaluators tend to avoid extreme scores, which can result in such imbalance in the dataset[Leckie and Baird, 2011]. Also, we can observe from Figure 3.13, that some algorithms are biased to lower grades, like LR, while others are biased towards higher grades, like SVR. This result demonstrates that the algorithm of an AES has effects on the score for the student, and then have significant consequences for the student.

Finally, our third research question is related to unstructured learning algorithms, like MLP. Although we can observe some variance between the algorithms and the aspects, it is clear that structured learning achieves superior performance. However, as we discuss in the next chapter, our datasets can contain what we call human bias in the scoring process. Next, we investigate if this kind of bias affects the results we obtained in this chapter.

We analyzed each skill according to the features we proposed as well. It is possible to observe that some features are more important than others when considering a skill, especially the features related to the formal aspect of the language. Yet, other skills are affected by the lack of specific features. The lack of features is due to the difficulty of finding Portuguese data labeled. For instance, it is possible to implement a more robust grammar error detector employing neural networks and then to use its results in the AES system. Unfortunately, there is no data for such training in the Portuguese language. Another possibility to leverage the results is the argumentation mining of the essays; however, this task requires labeling from multiple experts, and currently, there is only an English dataset for this purpose. Also, the Solution Proposal presents subjectivity nature, and sometimes requires a context outside of the essay. Then, there still much research to do regarding some skills.

In addition to that, we observed in these sections, the differences between evaluators employing a substantial amount of data. This statement is specific to our corpora since both presented a very similar distribution of features, as described in Section 3.2.1. Although the difference between human evaluations is a well-known phenomenon in education research, there is a lack of computational experiments in this subject. Therefore, our experiments endorse this idea and also point out when an evaluator may not consider the features that are expected for some competence.

Chapter 4

Automatic Essay Scoring in the Presence of Biased Ratings

It is well known that raters [Leckie and Baird, 2011] are subject to drift in scoring, in experience, and to the central tendency effect. In addition to that, Daniel Kahneman, Nobel prize in Economic Sciences in 2002, asserts in his book "Thinking, fast and slow "[Kahneman, 2011] that "bias judgments repeat in a predictable way when we are under specific circumstances." Therefore we supposed that during the human evaluation of an essay, there is an effect towards the scoring according to the evaluator's cultural background, political views, and attitudes.

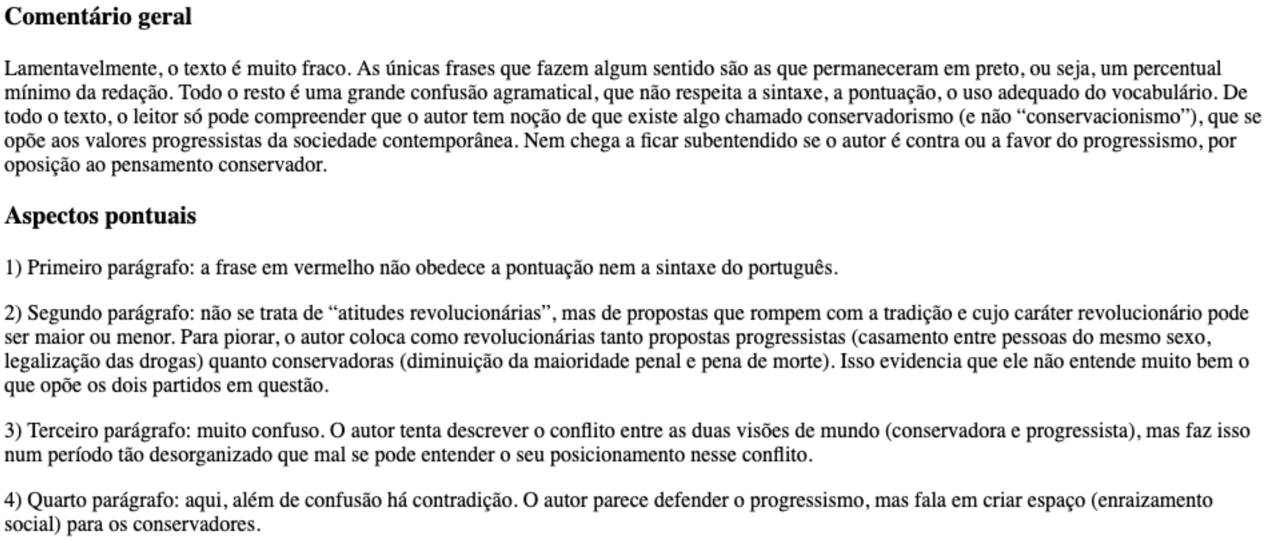
Considering this hypothesis, we aim to propose features that can indicate bias in the rater's comments. Also, we hypothesize that essays with comments that present strong bias has an impact on their scores. A deeper understanding of such an issue may help to mitigate the effects of rater bias, enabling AES models to achieve greater efficacy. From these hypotheses, we formulated the following research questions:

RQ4: Does subjectivity in rater comments vary depending on the given score?

RQ5: Does subjectivity in rater comments vary depending on the misalignment between the AES model and the human rater?

RQ6: Can we mitigate the effects of biased ratings?

To answer our research questions, we proposed a methodology that is composed of two main steps. In the first step, the features that describe subjectivity in the evaluator's comments are identified, then the second step tries to remove the biased scores from our training data. In the following sections, we describe these steps.



Comentário geral

Lamentavelmente, o texto é muito fraco. As únicas frases que fazem algum sentido são as que permaneceram em preto, ou seja, um percentual mínimo da redação. Todo o resto é uma grande confusão agramatical, que não respeita a sintaxe, a pontuação, o uso adequado do vocabulário. De todo o texto, o leitor só pode compreender que o autor tem noção de que existe algo chamado conservadorismo (e não "conservacionismo"), que se opõe aos valores progressistas da sociedade contemporânea. Nem chega a ficar subentendido se o autor é contra ou a favor do progressismo, por oposição ao pensamento conservador.

Aspectos pontuais

1) Primeiro parágrafo: a frase em vermelho não obedece a pontuação nem a sintaxe do português.

2) Segundo parágrafo: não se trata de "atitudes revolucionárias", mas de propostas que rompem com a tradição e cujo caráter revolucionário pode ser maior ou menor. Para piorar, o autor coloca como revolucionárias tanto propostas progressistas (casamento entre pessoas do mesmo sexo, legalização das drogas) quanto conservadoras (diminuição da maioridade penal e pena de morte). Isso evidencia que ele não entende muito bem o que opõe os dois partidos em questão.

3) Terceiro parágrafo: muito confuso. O autor tenta descrever o conflito entre as duas visões de mundo (conservadora e progressista), mas faz isso num período tão desorganizado que mal se pode entender o seu posicionamento nesse conflito.

4) Quarto parágrafo: aqui, além de confusão há contradição. O autor parece defender o progressismo, mas fala em criar espaço (enraizamento social) para os conservadores.

Figure 4.1. An example of a comment from a human evaluator in the UOL dataset.

4.1 Problem definition and Dataset

Comments written by human evaluators about essays can be subjective, and even show some bias against the stand of the writer. For instance, Figure 4.1 shows the comment of an evaluator about an essay from the UOL dataset. The last sentence of the comment says: "4) The fourth paragraph: here, besides the confusion, there is the contradiction. The author seems to defend the progressivism, but he says about making space (social rooting) for the conservatives". Although the evaluator details about the contradiction, it is unclear what are the confusions the writer states.

In this example, the evaluator only makes dubious claims. However, it is possible that in certain situations, the subjectivity of the evaluator reveals a stand for or against the student's opinion. Depending on the stand, the resulting score can be higher or lower. Nonetheless, it is desirable that the evaluations are as objective as possible and nobody is harmed or favored because of an opinion.

We aim to diminish this influence in the essay scores; then, we proposed a bias definition in the context of essay evaluation. Next, we define the level of the subjectivity of a document, and then what is a biased comment.

Definition 1. *Let L be a set of lexicons that indicates subjectivity in texts, and $d(x_1, x_2)$ a function of the distance between two texts x_1 and x_2. The level of the subjectivity of a given text t is the distance $d(L, t)$.*

Definition 2. *Let S be a set of essays, and $e \in S$ an essay such that its score is $g(e)$. Then, a comment about an essay e is biased when its level of subjectivity is in average bigger than in the following essays $S' = \{e' | \forall e' \in S, e' \neq e \wedge g(e') = g(e)\}$.*

Our definition of subjectivity degree is based on the work of Recasens et al. [2013], which used classes of subjectivity words to predict if a text is biased or not. However, our definition of biased text broadens the definition of Recasens et al. [2013]' work. This definition is also a contribution of this dissertation.

The following section details the instruments employed to extract from the evaluator's comments the features that describe bias.

4.2 Features for Identifying Biased Scores

Bias and subjectivity can be intrinsic to a domain, like political bias [Gentzkow and Shapiro, 2010], but some lexicons can characterize a stance towards a subject in different domains. In fact, bias and subjectivity for different types of scenarios are well documented in the linguistic literature. According to Verhagen [2005], the theory of enunciation accepts that the language allows that the speaker declares a stance implicitly, and some clues describe the subjectivity attached to the speaker's stance. The linguistic theory proposed by Anscombre and Ducrot [1983] states that language can reveal the stand of a person. The features that display the stand or bias of a person are composed of subjectivity elements such as argumentative markers, some specific kind of verbs, modals verbs, among other linguistic expressions. Recasens et al. [2013] employed some of these subjectivity elements in their work, like factive verbs, to detect the biased language in editions of Wikipedia.

Then based on the theories of enunciation and in the work of Recasens et al. [2013], Cançado et al. [2019] developed a handcraft list of Portuguese lexicons that indicate subjectivity in texts context-free, i.e., they appear in a multitude of types of text. The strategy to build this list comprises lexicons collected from comments of 50 essays and Portuguese lexicons based on the Koch [1984, 1992] books. First, the authors divided the words and expressions into two groups: the group of words and expressions that are meaningful and the sense of these words can be different according to the context, such words are usually names, verbs, and adjectives; and the group of words whose sense is context-independent, which usually are adverbs, prepositions, and conjunctions. Based on the linguistics theories and the intuition of Portuguese native speakers, the authors analyzed 522 collect texts, composed of 200 essays, 200 paper's abstract, and 122 newspaper columns opinion. Then, the result of this process is the lexicons that indicate bias.

The lexicons built were divided into the following five categories.

- **Argumentation**: This lexicon includes markers of argumentative discourse. Such kind of marker makes a degree of arguments, and also it introduces the strongest or the weakest argument. In addition to that, argumentative lexicons are the opinion footprint of the writer, which means that it can result in linguistic bias.

  Argumentative markers include lexical expressions and connectives, such as: "even" (até), "by the way" (alías), "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. *Ex.: a análise das causas é simplificadora e superficial, apesar de plausível* (the analysis of the causes is simplifying and superficial, though plausible).

  In the example, the evaluator makes a statement. Still, she introduces a counterpoint about her argument, which is contrary to the expectations of the reader. These lexicons link two opposite arguments; nonetheless, one of the arguments is more relevant than another, and that argument confirms the conclusion's writer.

- **Presupposition**: This lexicon includes markers that suggest the rater assumes something is true previously, which can be an opinion or a statement. Some examples of such markers include: "nowadays" (hoje em dia), "to keep on doing" (continuar a), and factive verbs. *Ex.: O que o autor parece não perceber é que o trote vem se tornando mais violento hoje em dia.* (What the author does not seem to realize is that student's prank is becoming more violent nowadays)

  In the example, the writer takes for granted that a student's prank is more common now than was before.

- **Modalization**: This lexicon indicates that the writer exhibits a stance towards its statement. According to the Logic, there are two kinds of modalizers, the possibilities ones, and the necessities ones. They can express the possibility and necessity regarding the belief of the writer about her statement, for instance, the term *É certo* (It is certain) expresses the sure that the speaker has about her claim. Also, they can manifest possibility or necessity regarding a rule, moral or legal. For example, *É obrigatório* (it requires) indicates an obligation about something.

Some other examples of such markers are adverbs, auxiliary verbs, modality clauses, and some type of verbs. *Ex.: Texto muito bom, apesar de desnecessariamente prolixo*. (Very good text, though unnecessarily wordy). In the example, the writer denotes a possibility regarding his opinion.

- **Opinion**: This lexicon also includes markers that indicate a state of mind or a sentiment of an opinion of the rater while evaluating the essay. For example, *É com grande satisfação* (It is with great satisfaction) is a term that reveals an opinion about an assertion, which can indicate the bias of the speaker.

  Some other examples of such markers include: "with regret" (infelizmente), "fortunately" (felizmente), and "it is preferable" (preferencialmente). *Ex.: Infelizmente, o texto está muito abaixo do que se espera do desempenho de um estudante ao final do Ensino Médio..* (Unfortunately, the text is far below what is expected of a student's performance at the end of high school.) In the example, besides the information about the performance, there is an opinion about the information.

- **Valuation**: This lexicon assigns a value to the facts, and it is represented by the states or qualities ascribed to a subject. Adjectives are usual to assess the value of a subject and also to reflect the stance of the speaker. Therefore, it is a cue to detect bias in the writer's speech. However, as adjectives are context-dependent, we use only in this class the markers that are related to intensification, which indicate opinion in the speech as well. *demais* (too much), is an example of an intensifier of a fact, probably exaggerating the statement.

  Other examples of intensifiers are: "absolutely" (absolutamente), "highly" (altamente), and "approximately" (aproximadamente). *Ex.: O que a existência de oposição tem a ver com a intolerância religiosa não fica absolutamente claro.* (What the existence of opposition has to do with religious intolerance is not absolutely clear). In the example, the subjective attitude of the evaluator is measured by the valuation of the statement.

All the examples listed above were excerpts from the essay's comments to provide the intuition about the selection of such words. The list of the lexicons are free for download[1] and also are listed in Appendix F.

[1] url to download

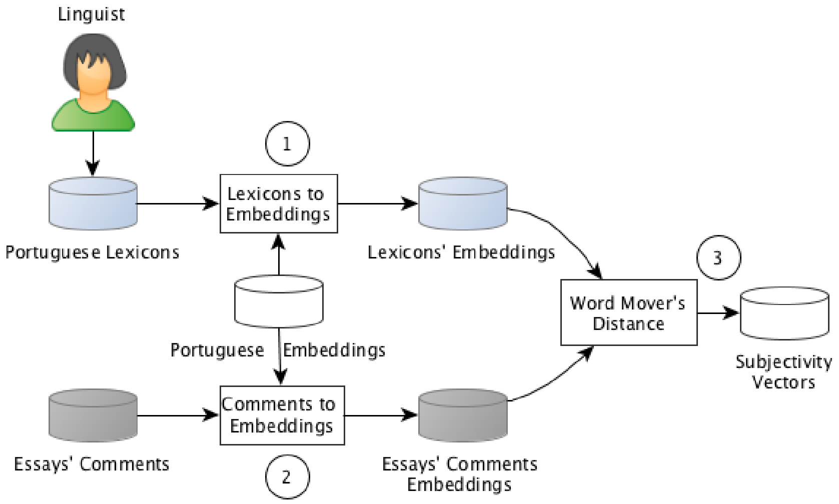


Figure 4.2. The steps (1) and (2) for the debiasing process

4.3 Debiasing the Training Set

According to Kahneman [2011], biases are "systematic errors" that are repetitive and predictable in specific scenarios. Still, an objective observer is prone to detect such errors that we make. Usually, errors are events that are different from what is expected, which is commonly the average of the population. Then we suppose that if subjectivity is a metric for biased language, and the subjectivity level of a comment of an essay diverge of the average, then we assume that this is an error in this scenario and is related to the bias, which can favor or disfavor the student's grade. Therefore we define that the detection of biased comments of the human evaluator is to verify if the subjectivity of the evaluator's comments is close of the subjectivity norm, which is computed using the comments that have essays scored with the same value.

The steps for the debiasing the dataset are the following: (1) computation of word embeddings; (2) the building of the subjectivity vectors; (3) computation of centroids for each score; (4) Removal of the α essays most distant of its centroid. Figure 4.2 depicts the steps (1) and (2) and Figure 4.3 depicts steps (3) and (4).

Next, each step is described in detail.

Computation of word embeddings: Word embeddings [Mikolov et al., 2013a] are the representation of words as dense vectors. The methodology that is employed to build such representations are described in Section 2.1.2, and a scheme is depicted in Figure 2.2.

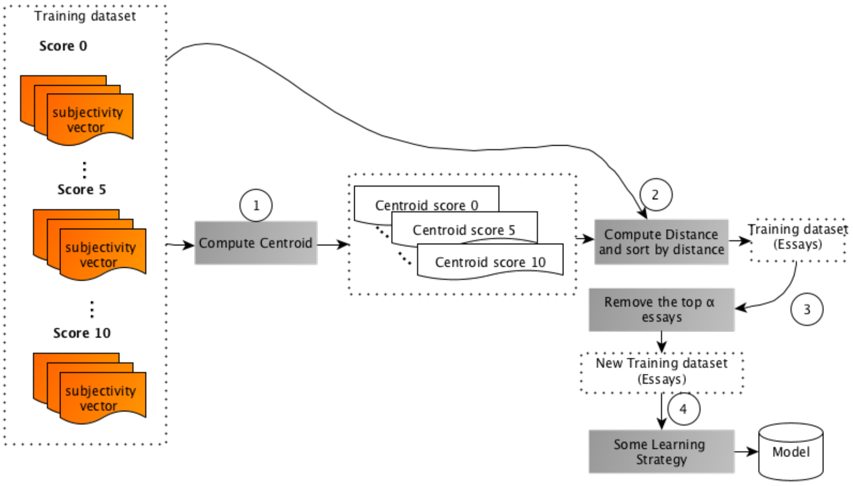


Figure 4.3. The steps (3) and (4) for the debiasing process

The building of the subjectivity vectors: A subjectivity vector is a five dimension vector, each one representing a subjectivity category as described in Section 4.2. Also, we consider that each subjectivity category as a text document composed by its set of lexicons, then the category is transformed into a dense vector employing the embeddings produced in the previous step. Next, the evaluator's comment is also transformed into a dense vector, and we compute the distances between this vector and the dense vectors of each subjectivity category using Word Mover's Distance algorithm [Kusner et al., 2015]. Each distance computed results in a dimension of the subjectivity vector.

Computation of centroids for each score: For each score, we take their subjectivity vectors of the essay's comments, and compute the centroid. The centroids of the subjectivity vectors are a kind of prototype of a score group.

Removal of the α essays most distant of its centroid: Being α an input variable and a percentage, we remove α essays that are the most distant from their centroids. If we remove α essays from the training set, then these essays have comments whose subjectivity is very different from their prototype. As it is unfeasible to know beforehand the rate of biased comments, we choose to test several α values and compare the results.

4.4 Experiments

Our experiments are composed of the subjectivity analysis of our dataset and the study of the influence of the subjectivity in the human scores. First, we inspect the subjectivity vectors of the evaluator's comments and their relation with the scores of their essays. This experiment motivates us to perform a second experiment, the analysis of essays with anomalous comments regarding its subjectivity vector.

To perform such experiments, we employ the first version of the UOL dataset of 1,840 essays. Also, from these essays, we separated some essays ($n = 50$), which received scores by two expert raters who were directly instructed to perform impartially, objective, and unbiased evaluations. These raters are PhD-level in Linguistics with unlimited time to provide their ratings, and they do not participate in the creation of the training set. We assume biased judgments did not contaminate the ratings given to these essays. We used these essays to evaluate the efficacy of AES models learned after the training set is debiased.

Analysis of Subjectivity Vectors: This experiment is concerned with RQ4, which aims to reflect on the relation between the subjectivity and the scores. In our first analysis, we employed human evaluation. Figure 4.4 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 4.4, 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.

In addition to that, we also analyzed how the centroids are distant from each other. Figure 4.5 shows t-SNE representations [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 7 and 10.

Misalignment between scores: Our next experiment is concerned with RQ5. First, we define as misalignment as the difference between the score assigned by a human evaluator and the score assigned by an algorithm. Figure 4.6 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

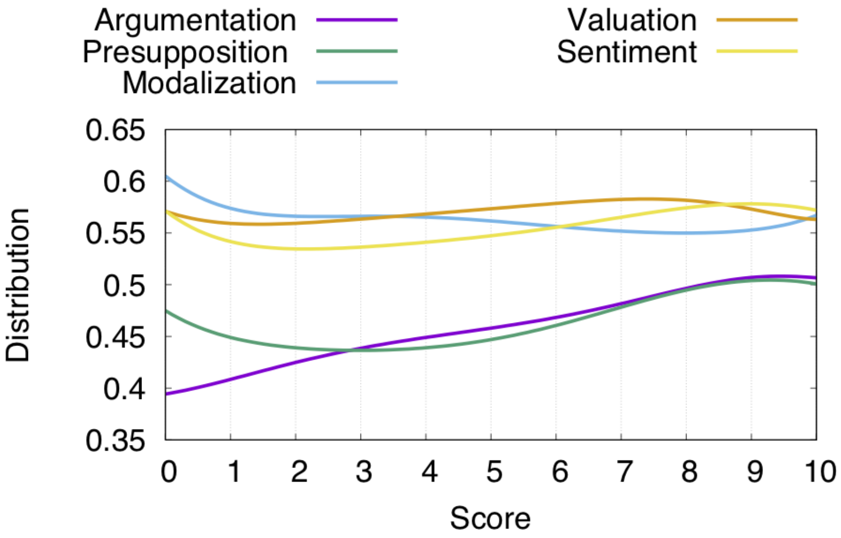


Figure 4.4. Subjectivity distribution for human raters.

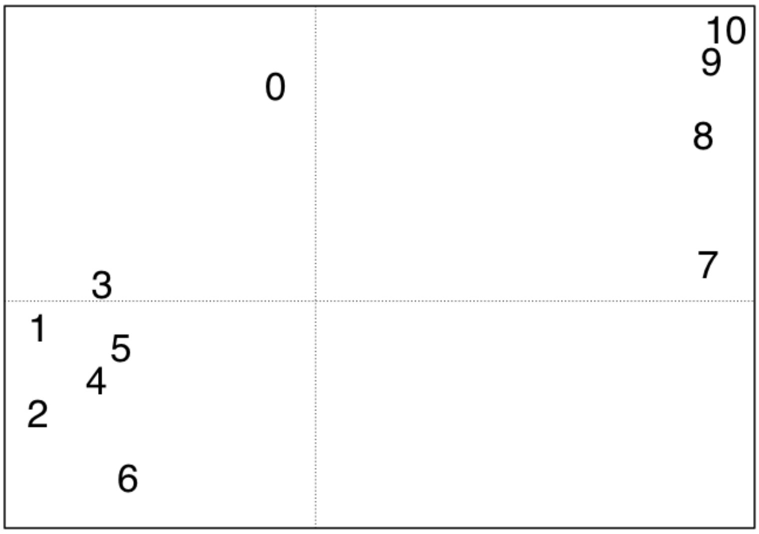


Figure 4.5. t-SNE representation for subjectivity vectors. Numbers correspond to the scores assigned to corresponding essays.

the human rater. Then, we calculated the average subjectivity vector for each group. As we can see, subjectivity affects AES models in different ways. In general, however, subjectivity vectors within groups of essays associated with extreme misalignment are very different from subjectivity vectors associated with mild misalignment.

Figure 4.7 shows t-SNE representations for subjectivity vectors grouped by mis-

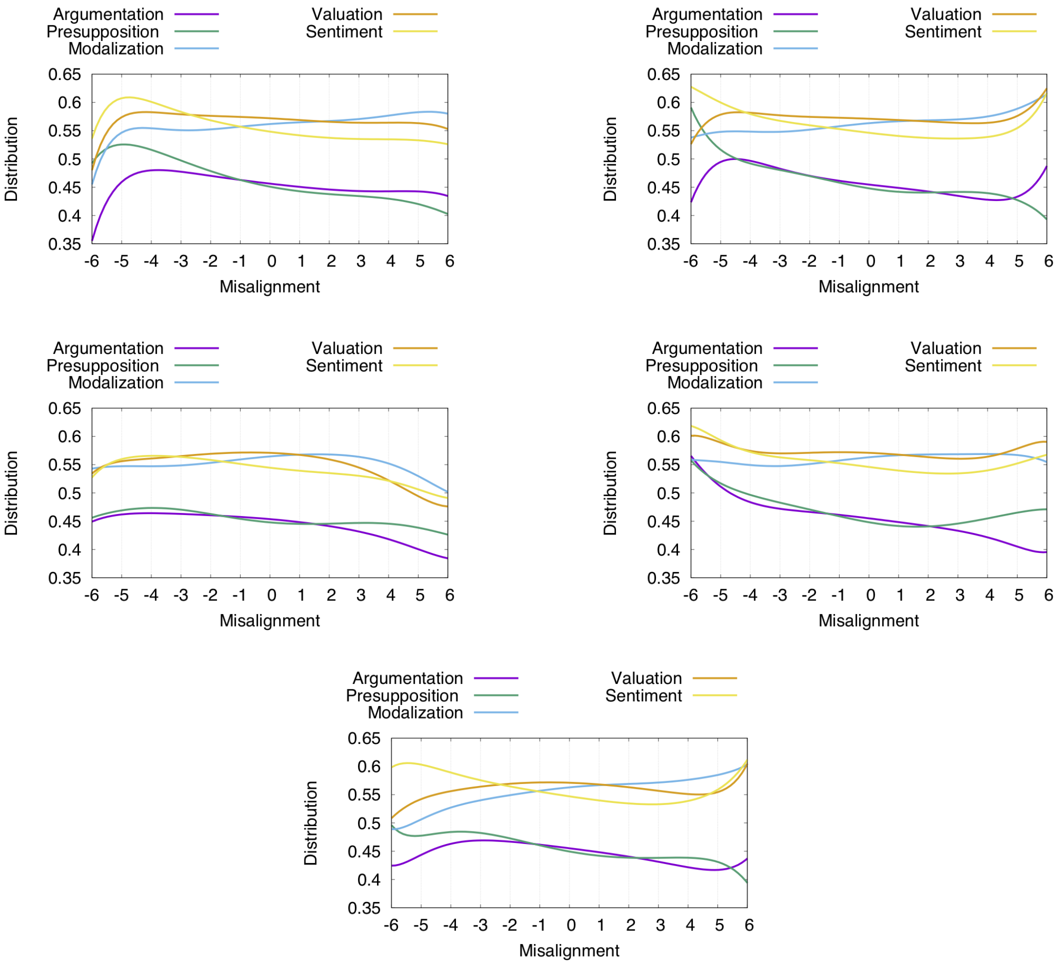


Figure 4.6. Subjectivity distribution. (Top to bottom) SVR, RF, LR, GB, and MLP.

alignment levels. Each cluster contains ≈ 80% of the vectors associated with one of the misalignment levels inside the cluster. That is, 20% of the essays are removed from the training set (i.e., $\alpha = 0.2$).

Debiasing the training set: The last experiment is concerned with RQ6. As described earlier, our debiasing approach works by removing from the training set some essays (controlled by α) that are more likely to be associated with biased ratings. Table 4.1 shows κ numbers for different α values. 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.

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

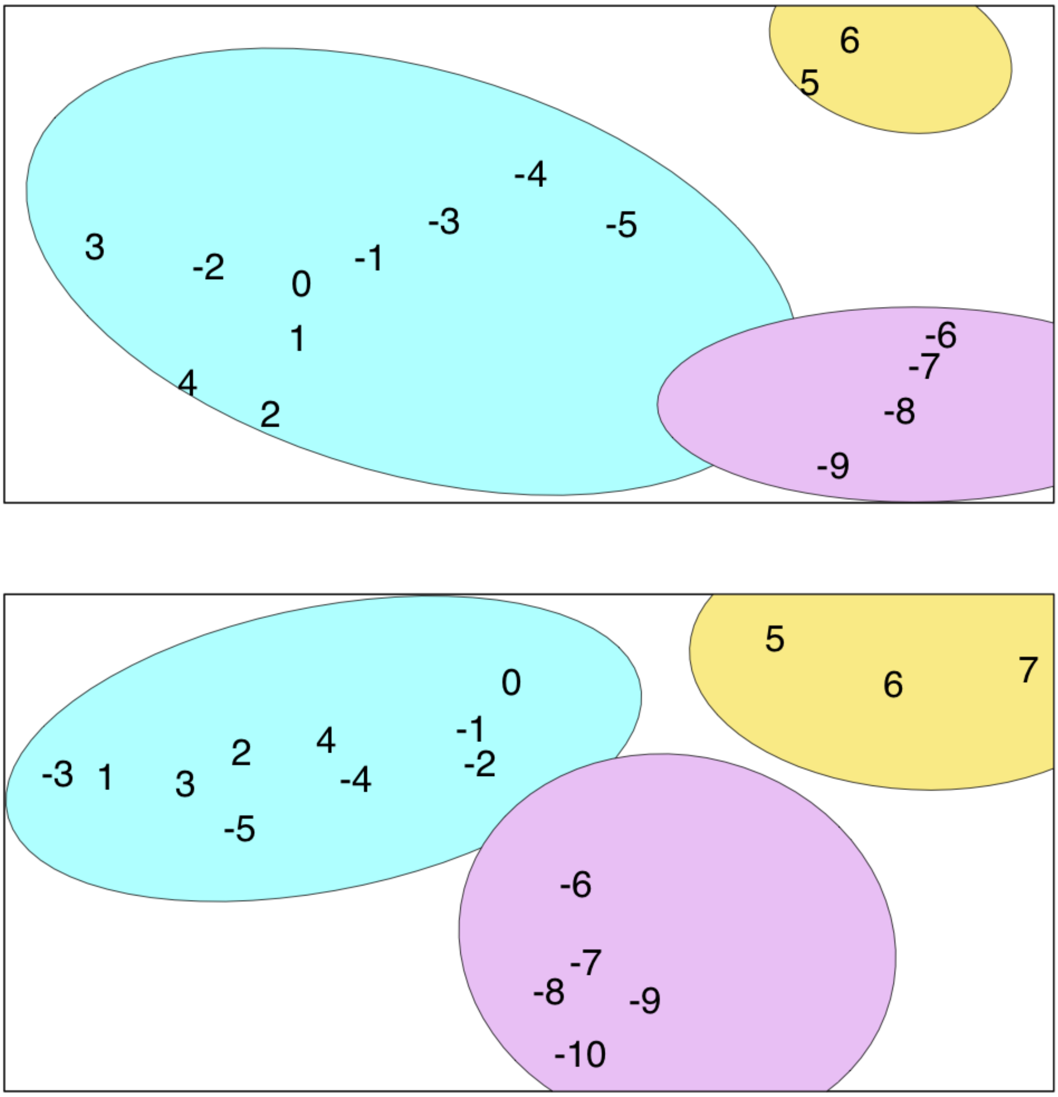


Figure 4.7. t-SNE representation for subjectivity vectors grouped by misalignment levels. The corresponding regions comprise essays associated with specific misalignment levels. (Top) GB model. (Bottom) MLP model.

the training set, and the test set is composed of unbiased ratings. Table 4.2 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. The increase in kappa is either because we started to remove unbiased ratings, or because the training set became too small. In all cases, the MLP model showed to be statistically superior to the other models.

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

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

4.5 Discussion

Human unconscious bias affects the scoring of argumentative essays, considering such a scenario, we tried to model the human bias employing subjectivity lexicons divided into five categories. Each category represents a dimension of the bias concept, then the bias of a text document is depicted by a vector with these five dimensions. To build this vector, we employed embeddings and WMD, as described in Section 4.3. We called this structure the subjectivity vector.

The use of the subjectivity vector allowed us to analyze the subjectivity in human comments. By conducting an investigation that compared subjectivity vectors of different scores, we observed that the subjectivity of score present differences. Figure 4.5 shows that the centroids are separated. And although we can cluster the centroids in three groups, Figure 4.4 makes even more explicit the differences between the scores and each subjectivity dimension.

In addition to that, we showed that the higher the misalignment between human score and machine score, then we have greater subjectivity levels. Figure 4.7 depicts this idea in an understandable way for GB and MLP models. Also, each algorithm was evaluated according to the subjectivity, and Figure 4.6 reports such evaluation.

Finally, we examined the influence on the results of the algorithms. A set of $n = 50$ essays, which were set apart from the main dataset, was evaluated by experts as biased or not. Table 4.1 shows that the bias occurs in our dataset, and the efficacy of our models decreases. Table 4.2 presents a test using our debiasing approach with the whole dataset $n = 1,840$, and the results are similar to the smaller dataset.

Thus, we were able to answer our research questions through a novel methodology and several experiments. It is possible to present more questions regarding this rich dataset and perform more experiments. Therefore in the next chapter, we discuss the results of this dissertation and also present new questions to answer in future work.

Chapter 5

Conclusion

In this dissertation, we presented a system whose final goal is to score essays in fairly. There is still some work to do to finish this thesis, and in the next sections, we summarize our results until now, and also we present some suggestions for future work.

5.1 Main Results

The automatic essay scoring (AES) is a task that has already been studied by the Natural Language processing researchers for several years. In this dissertation, we approached AES in a new looking way. First, we investigated two Portuguese datasets regarding five scoring aspects and the explainability of the models; then, we analyzed the comments of the human evaluators, and how biased comments can affect the efficacy of our models. These two phases of our study were based on the following research questions:

RQ1: Which are the most relevant features to a structured technique like the Logistic Regression (LR)? Can we have some insight into human evaluation regarding the aspects of essays?

RQ2: How scores are distributed across the essays? How aligned with human raters are different AES models?

RQ3: Structured and Unstructured present comparable performance in multi-aspect Automatic Essay Scoring task?

RQ4: Does subjectivity in rater comments vary depending on the given score?

RQ5: Does subjectivity in rater comments vary depending on the misalignment between the AES model and the human rater?

RQ6: Can we mitigate the effects of biased ratings?

In Chapter 3, we discussed the questions RQ1, RQ2, and RQ3. Such examination began with an LR experiment, then using an ablation test of features, we measured the influence of the features in each scoring aspect. This experiment led to interesting insights. First, we confirmed results from several Educational works that different evaluators weights rubrics distinctly. An evaluator can even assess rubric disregarding features related to that rubric. Moreover, there are some features, like the number of different words that influence all aspects and the final score. Other features are like conclusion markers, that are only relevant to some specific rubric. Also, we performed experiments with a variety of algorithms, namely: XGboost (GB), Random Forest (RF), Multi-Layer Perceptron (MLP), and Linear Regression (LR). An additional analysis that we experimented with was the SHAP [Lundberg and Lee, 2017] strategy, which was employed to understand to what extent the features influence each aspect.

Another experiment that we conducted was an analysis of the misalignment between humans and algorithms. By misalignment, we understand the difference between the score assigned by the human evaluator and a given algorithm. Figure 3.14 shows explicitly how some methods are more strict than others, i.e., some methods predict a lower score than humans, while others predict a higher score than humans. Finally, we examined the results of all algorithms since each algorithm employs a different strategy to predict the score, and the conclusion is that MLP presented the most robust results for our features.

In the next chapter, we considered the questions RQ4, RQ5, and RQ6. To answer these questions, a linguist built a lexicon list, which comprises five classes of terms that are subjective for different types of texts, like news, social media comments, and scientific papers. Each class in this list is represented as word embeddings, and then we compute the distance between each embeddings' class and a given evaluator's comment. The distances calculated are stored in a vector called subjectivity vector that represents the subjectivity of the given evaluator's comment. Using the subjectivity vectors, we first analyzed the subjectivity level for humans evaluators and the algorithms tested in Chapter 3. As subjectivity level for a given score, we consider as the centroid of the subjectivity vectors for that score. The results are depicted in Figure 4.4 and Figure 4.5. In addition to that, we also proved that the subjectivity level varies according to the misalignment (Figure 4.6).

Another issue discussed in this dissertation is how a biased evaluation impacts an AES. Thinking about this issue, we built a small essay dataset whose evaluations were labeled as biased or not. Our experiments (Table 4.1) showed indications of a negative impact on the models' efficacy. Thus we suppose that in the whole dataset, the algorithms would behave in the same way, then we performed several experiments removing gradually a percentage of essays with highly subjectivity comments. Again we obtained evidence that suggests a negative impact of the bias in the models' efficacy. Probably this happens because even though the essay presents features for a given score, the human label confuses the classifier, which then misleads it.

5.2 Discussion

In the introduction of this dissertation, we listed our main contributions. Four of them are related to Chapter 3 and are as follows.

- we built open corpora of labeled Portuguese essays (Section 3.1);
- we proposed an Automatic Essay Scoring (AES) framework for the Portuguese language (Section 3.2);
- we proposed a Multi-aspect essay scoring system (Section 3.2);
- we analyzed in detail the relevant features for our Portuguese AES framework for each aspect (Section 3.3.1);

We proved that we accomplished these contributions even with the limited amount of Portuguese tools of NLP. In Section 3.2, we proposed several handcrafted features for our system. Considering the nature of our task and the attributes of our corpora, we classified our features as baseline features and domain features. The baseline features were based on previous work about AES [Attali and Burstein, 2006], and they resulted in a robust baseline system. On the other hand, the domain features were based on the rubrics of the *Exame Nacional do Ensino Médio* (ENEM). Tables 3.5 and 3.6 describe the outcomes of our experiments in the proposed corpora. Thus, we showed our contribution as the first AES system for the Portuguese language. This contribution is notably meaningful since, in Brazil, there is the ENEM exam, which has millions of people enrolled every year[1]. This contribution leads us to the other two contributions. One is the discussion about the adoption of an AES in one of the biggest

[1] `http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/enem-2018-tem-6-7-milhoes-de-inscritos/21206`

tests in the world. The use in the large scale of an AES system in ENEM would help to reduce government costs, besides the decrease of human errors like tiredness. The diminishing of human errors in ENEM is crucial to Brazilian society since the ENEM is employed as an entrance exam to most of Brazilians colleges and universities. Another contribution is to help the debate about the standardization of education in Brazil, which lacks a national standard for its educational system. As Brazil does not adopt a standard to its educational system, students all around the country learn differently, and then low-income students are harmed by poor quality teaching. In standardize strategy, all students would learn similar content, hence decreasing education inequality. The discussion of this subject is essential to the development of Brazil.

Besides an automatic essay scoring system, we proposed to score the aspects of an essay. These aspects are based on the skills of the ENEM exam. As far as we know, none AES system evaluates these aspects. Thus, as our proposal is unique, this is another contribution. Also, we assessed the features related to each aspect. This analysis demonstrated that each human evaluator assesses features differently. This result corroborates qualitative results from Education works[Freedman, 1979]. The feature analysis is, therefore, another contribution since we proved quantitatively such results that were only demonstrated in qualitatively. Such a conclusion is fortified by the similarity of our dataset (Section 3.2.1). Finally, it is the first time that open corpora of Portuguese essays are built. This contribution aids the NLP community to leverage the results in the AES task and related NLP tasks as well.

The other contributions of this dissertation are related to the Chapter **??**. Next, we listed these contributions.

- we proposed a categorization of bias to detect in a text since this is a word with several meanings (Section 4.1);
- we introduced the definition of the degree of subjectivity of a text, and from this, we proposed the definition of biased text (Section 4.1).
- we introduced a new unsupervised method to detect bias in texts (Section 4.3).
- we proved that biased labeling harms the classification results (Section 4.4).

Our first contribution is the categorization of bias, which has several meanings. The categorization helps the research about the topic, and then leverage the results in the are. We also introduced the definition of degree of subjectivity in a text. This definition is groundbreaking because it considers lexicons as cues to a biased text. Although our definition is based on the work of Recasens et al. [2013], our features are distances between dense vectors of lexicons and dense vectors of a text. Recasens et al. [2013] handcrafted features based on their lexicons. We used this definition to establish what is a biased text, which is undefined in all works that we found. From the biased text definition, we proposed a new approach to detect implicit bias in a text. We proved that our approach detects the human bias in evaluations of essays. Pieces of evidence are described in the results of the experiments depicted in Tables 4.1 and 4.2. Finally, we demonstrated that the bias in the labeling harms the prediction of a machine learning algorithm (Section 4.4). These contributions raise crucial questions that should be debated and also can see as contributions. One question is the essay to be the primary selection method to the entrance in the university. Would it be a subjective method like essays an adequate selection to college entrance? Is it possible to change the way the essay is employed as a selection method? Another issue is the influence of the life experience of the human evaluator in assessing essays. Is it possible that a high load of training of the human evaluators improve such influence? Is it worth to invest in training for the human evaluators? All these questions are hard to answer; however, the debate about these and other related questions should be done.

This dissertation instigates an intense discussion about several issues. Besides the contributions, new views and reflections were promoted. Thus, in the next section, we suggest future works based on the contributions, views, and reflections discussed in this section.

5.3 Future Work

We achieved some results and conclusions for this proposal, and as the outcome, we published the following two papers:

- Amorim, E., & Veloso, A. (2017). A Multi-aspect Analysis of Automatic Essay Scoring for Brazilian Portuguese. In Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics (pp. 94-102).

- Amorim, E., Cançado, M., & Veloso, A. (2018). Automated Essay Scoring in the Presence of Biased Ratings. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (Vol. 1, pp. 229-237).

Also, we submitted a paper to the Lingumática Journal[2].

[2] https://linguamatica.com/index.php/linguamatica

However, we believe that there is much more to investigate in this subject, we suggest as future work the following topics:

1. **Prediction of Comments**: A fascinating investigation is to predict the score of essays, but instead of using the text of essays, we would use the text of the evaluator's comments about the given essay. The comments that are misclassified could be analyzed employing the explicability tools of neural networks.

2. **Improvements of auxiliary tools**: As we explained in Chapter 3, due to the lack of robust NLP tools for Portuguese, the performance of an AES system is harmed. An efficient grammar error detection system is the cornerstone of a robust AES strategy; then, we suggest that as future work, the development of such kind of tool to leverage our results.

3. **Transfer Learning of Essay's topics**: Although the tests of this dissertation comprise essays from several topics, usually AES experiments are performed inside the same topic. However, it is not always possible to collect enough essays on the same topic. Then, works that investigate the transfer learning between essay's topics would be instrumental.

4. **Experiments in an English dataset**: Also, we intend to experiment with our current strategy and new strategies in a dataset in English. There are a lot of English essay datasets; then, it is straightforward to experiment with our strategies in one of these datasets. We are already implementing the handcraft AES system and the NN AES system for the English Language, and we expect to produce results soon.

Bibliography

Alikaniotis, D., Yannakoudakis, H., and Rei, M. (2016). Automatic Text Scoring Using Neural Networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 715--725.

Amorim, E. and Veloso, A. (2017). A multi-aspect analysis of automatic essay scoring for brazilian portuguese. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94--102.

Anscombre, J.-C. and Ducrot, O. (1983). *L'argumentation dans la langue.* Editions Mardaga.

Attali, Y. and Burstein, J. (2006). Automated essay scoring with e-rater V. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1--34.

Breland, H. M., Jones, R. J., Jenkins, L., Paynter, M., Pollack, J., and Fong, Y. F. (1994). The college board vocabulary study. *ETS Research Report Series*, 1994(1).

Brenner, H. and Kliebsch, U. (1996). Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*, pages 199--202.

Bridgeman, B. (2013). Human ratings and automated essay evaluation. *Handbook of automated essay evaluation: Current applications and new directions*, pages 221--232.

Bryant, C., Felice, M., Andersen, Ø. E., and Briscoe, T. (2019). The bea-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52--75.

Bryant, C. and Ng, H. T. (2015). How far are we from fully automatic high quality grammatical error correction? In *ACL (1)*, pages 697--707.

Burstein, J., Chodorow, M., and Leacock, C. (2003). Criterionsm online essay evaluation: An application for automated evaluation of student essays. In *IAAI*, pages 3--10.

Burstein, J., Marcu, D., Andreyev, S., and Chodorow, M. (2001). Towards automatic classification of discourse elements in essays. In *Proceedings of the 39th annual Meeting on Association for Computational Linguistics*, pages 98--105. Association for Computational Linguistics.

Cançado, M., Amaral, L. L., Mello, H., Amorim, E., and Veloso, A. (2019). Detecção automática de viés linguístico em correções de redação. Manuscript.

Changhuo, X., Dong, C., Qian, W., and Zhilan, X. (2015). On automated essay scoring for learners of chinese as a second language. *International Chinese Teaching and Research*, 1:015.

Chen, H. and He, B. (2013). Automated essay scoring by maximizing human-machine agreement. In *EMNLP*, pages 1741--1752.

CoGrOO, C. (2012). *CoGrOO: Corretor Gramatical acoplável ao LibreOffice e Apache OpenOffice*. CCSL IME/USP, São Paulo, Brasil.

Cozma, M., Butnaru, A. M., and Ionescu, R. T. (2018). Automated essay scoring with string kernels and word embeddings. *arXiv preprint arXiv:1804.07954*.

Dale, R. and Kilgarriff, A. (2011). Helping our own: The hoo 2011 pilot shared task. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 242--249. Association for Computational Linguistics.

Dodge, Y. (2008). *The concise encyclopedia of statistics*. Springer Science & Business Media.

Dong, F. and Zhang, Y. (2016). Automatic Features for Essay Scoring ï¿½ An Empirical Study. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing(EMNLP2016)*, 1966:1072--1077.

Dong, F., Zhang, Y., and Yang, J. (2017). 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.

Faulkner, A. (2014). Automated classification of stance in student essays: An approach using stance target information and the wikipedia link-based measure. *Science*, 376(12):86.

Fonseca, E., Medeiros, I., Kamikawachi, D., and Bokan, A. (2018). Automatically grading brazilian student essays. In *International Conference on Computational Processing of the Portuguese Language*, pages 170--179. Springer.

Freedman, S. W. (1979). How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology*, 71(3):328.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

Gentzkow, M. and Shapiro, J. M. (2010). What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35--71.

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345--420.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Habernal, I. and Gurevych, I. (2016). What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223.

Hancke, J. and Meurers, D. (2013). Exploring cefr classification for german based on rich linguistic modeling. *Learner Corpus Research*, pages 54--56.

Higgins, D., Burstein, J., Marcu, D., and Gentile, C. (2004). Evaluating multiple aspects of coherence in student essays. In *HLT-NAACL*, pages 185--192.

Iyyer, M., Enns, P., Boyd-Graber, J., and Resnik, P. (2014). Political ideology detection using recursive neural networks. In *Proceedings of the Association for Computational Linguistics*, pages 1--11.

Jubran, C. C. A. S. and Koch, I. G. V. (2006). *Gramática do português culto falado no Brasil: construção do texto falado*, volume 1. UNICAMP.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kakkonen, T. and Sutinen, E. (2004). Automatic assessment of the content of essays based on course materials. In *Information Technology: Research and Education, 2004. ITRE 2004. 2nd International Conference on*, pages 126--130. IEEE.

Ke, Z., Carlile, W., Gurrapadi, N., and Ng, V. (2018). Learning to give feedback: Modeling attributes affecting argument persuasiveness in student essays. In *IJCAI*, pages 4130--4136.

Koch, I. G. V. (1984). *Argumentação e linguagem.* Cortez Editora.

Koch, I. G. V. (1992). *A inter-ação pela linguagem.* Contexto.

Koch, I. G. V. (1999). *A coesão textual.* Editora Contexto.

Koehn, P. (2009). *Statistical machine translation.* Cambridge University Press.

Kusner, M. J., Sun, Y., Kolkin, N. I., and Weinberger, K. Q. (2015). From word embeddings to document distances. In *ICML*.

Larkey, L. S. (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*, pages 90--95.

Leckie, G. and Baird, J.-A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, 48(4):399--418.

Li, J. and Hovy, E. H. (2014). A model of coherence based on distributed sentence representation. In *EMNLP*, pages 2039--2048.

Li, J. and Jurafsky, D. (2016). Neural net models for open-domain discourse coherence. *arXiv preprint arXiv:1606.01545*.

Lin, Z., Ng, H. T., and Kan, M.-Y. (2011). Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 997--1006. Association for Computational Linguistics.

Liu, H., Ye, Y., and Wu, M. (2018). Ensemble learning on scoring student essay. In *2018 International Conference on Management and Education, Humanities and Social Sciences (MEHSS 2018)*. Atlantis Press.

Lottridge, S. M., Schulz, E. M., and Mitzel, H. C. (2013). Using automated scoring to monitor reader performance and detect reader drift in essay scoring. *Handbook of Automated Essay Evaluation: Current Applications and New Directions, MD Shermis and J. Burstein, Eds. New York: Routledge*, pages 233--250.

Lukin, S., Anand, P., Walker, M., and Whittaker, S. (2017). Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, pages 742–753.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765--4774. Curran Associates, Inc.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579--2605.

Madnani, N., Heilman, M., Tetreault, J., and Chodorow, M. (2012). Identifying high-level organizational elements in argumentative discourse. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 20--28. Association for Computational Linguistics.

Manning, C. D., Raghavan, P., Schütze, H., et al. (2008). *Introduction to information retrieval*. Cambridge university press Cambridge.

Martins, E. (2000). *Manual de redação e estilo*. O Estado de São Paulo.

Martins, T. B., Ghiraldelo, C. M., Nunes, M. d. G. V., and de Oliveira Junior, O. N. (1996). *Readability formulas applied to textbooks in brazilian portuguese*. Icmsc-Usp.

Mendes, E. A. d. M. (2013). A avaliação da produção textual nos vestibulares e outros concursos: a questão da subjetividade. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, 18(2):435--458. ISSN 1414-4077.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111--3119.

Napoles, C. and Callison-Burch, C. (2015). Automatically scoring freshman writing: A preliminary investigation. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 254--263.

Ng, H. T., Tetreault, J., Wu, S. M., Wu, Y., and Hadiwinoto, C. (2013). The conll-2013 shared task on grammatical error correction. In *CoNLL Shared Task*, pages 1--12.

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., and Bryant, C. (2014). The conll-2014 shared task on grammatical error correction. In *CoNLL Shared Task*, pages 1--14.

Östling, R., Smolentzov, A., Tyrefors Hinnerich, B., and Höglin, E. (2013). Automated essay scoring for swedish. In *The 8th Workshop on Innovative Use of NLP for Building Educational Applications, Atlanta, GA, USA, June 13, 2013*, pages 42--47. Association for Computational Linguistics.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *The Journal of experimental education*, 62(2):127--142.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825--2830.

Perini, M. A. (2017). *Gramática descritiva do português brasileiro*. Editora Vozes Limitada.

Persing, I., Davis, A., and Ng, V. (2010). Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229--239. Association for Computational Linguistics.

Persing, I. and Ng, V. (2014). Modeling prompt adherence in student essays. In *ACL (1)*, pages 1534--1543.

Persing, I. and Ng, V. (2015). Modeling argument strength in student essays. In *ACL (1)*, pages 543--552.

Recasens, M., Danescu-Niculescu-Mizil, C., and Jurafsky, D. (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.

Silva Neto, S. R. d. et al. (2017). Uma abordagem computacional para identificação de indício de preconceito em textos baseada em análise de sentimentos.

Søgaard, A., Plank, B., and Hovy, D. (2014). Selection bias, label bias, and bias in ground truth. In *COLING (Tutorials)*, pages 11--13.

Song, W., Fu, R., Liu, L., and Liu, T. (2015). Discourse element identification in student essays based on global and local cohesion. In *EMNLP*, pages 2255--2261.

Stab, C. and Gurevych, I. (2014). Identifying argumentative discourse structures in persuasive essays. In *EMNLP*, pages 46--56.

Stab, C. and Gurevych, I. (2016). Parsing argumentation structures in persuasive essays. *arXiv preprint arXiv:1604.07370*.

Tay, Y., Phan, M. C., Tuan, L. A., and Hui, S. C. (2018). Skipflow: incorporating neural coherence features for end-to-end automatic text scoring. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Vajjala, S. and Loo, K. (2013). Role of morpho-syntactic features in estonian proficiency classification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 63--72.

Verhagen, A. (2005). *Constructions of intersubjectivity: Discourse, syntax, and cognition*. Oxford University Press on Demand.

Wikipedia (2017). Wikipedia: Neutral point of view. `https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view`. Accessed: 2017-10-01.

Yano, T., Resnik, P., and Smith, N. A. (2010). Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 152--158. Association for Computational Linguistics.

Yuan, Z., Briscoe, T., and Felice, M. (2016). Candidate re-ranking for smt-based grammatical error correction. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 256--266.

Zafar, M. B., Gummadi, K. P., and Danescu-Niculescu-Mizil, C. (2016). Message impartiality in social media discussions. In *ICWSM*, pages 466--475.

Zesch, T., Wojatzki, M., and Scholten-Akoun, D. (2015). Task-independent features for automated essay grading. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 224--232.

Zhou, Y. and Cristea, A. I. (2016). Towards detection of influential sentences affecting reputation in wikipedia. In *Proceedings of the 8th ACM Conference on Web Science*, pages 244--248. ACM.

Appendix A

Rules of Language Tool

Each rule in the LanguageTool software capture a string pattern, then to summarize all the patterns, 129 patterns in total, we get only the messages associated with several patterns. Therefore, a message can be associated to more than one pattern.

1. Não ocorre crase antes de palavras masculinas.
2. Não há crase neste caso, somente no plural ("‘as").
3. Não acontece crase antes de verbo.
4. Ocorre crase em express oes indicativas de horas.
5. Ocorre crase em express oes indicativas de horas.
6. Ocorre crase em express oes indicativas de modo, tempo, lugar etc.
7. A expressão "em relação" rege a preposição "a". Se for seguida de substantivo feminino singular determinado, haverá crase.
8. "Em relação" rege a preposição "a", logo há crase aqui.
9. "Com relação" rege a preposição "a", logo há crase aqui.
10. "Devido" rege a preposição "a". Se for seguido de substantivo feminino singular determinado, haverá crase.
11. "devido" rege a preposição "a". Se for seguido de substantivo feminino plural determinado, haverá crase.

12. Pronomes de tratamento não admitem artigo, portanto não haverá crase antes deles. A única exceção é o pronome "senhora", que admite artigo e, se houver também preposição, haverá crase.

13. O adjetivo concorda em gênero (masculino ou feminino) e número (singular ou plural) com o substantivo a que se refere.

14. A expressão "em anexo" é invariável.

15. O adjetivo "anexo" é variável, portanto concorda com o substantivo a que se refere.

16. A palavra "meio" usada no sentido de "um pouco" é advérbio, portanto invariável. Exceção: meia/meias pode ser substantivo. Exemplo: Minhas meias azuis estão manchadas.

17. A palavra "meio" usada no sentido de "um pouco" é advérbio, portanto invariável.

18. A palavra "meio" usada no sentido de metade é adjetivo numeral fracionário, portanto concorda com o substantivo a que se refere.

19. O verbo fazer, quando indica tempo, é impessoal e deve permanecer sempre no singular.

20. Na escolha entre "a" e "há", sempre que indicar tempo decorrido opte por "há", que corresponde ao verbo haver, em forma impessoal, sempre no singular.

21. As formas do verbo "haver" ficam no singular quando indicam tempo decorrido.

22. O verbo haver no sentido de existir é impessoal. Permanece sempre na terceira pessoa do singular.

23. Se o pronome "mim" é sujeito do verbo no infinitivo, o pronome a ser usado é "eu".

24. O pronome "eu" não pode ser regido de preposição. (Neste caso a preposição é "entre".)

25. Mau é adjetivo (o feminino é "má" e o plural é "maus") e mal é advérbio (forma invariável). Para distinguir o uso adequado de mal/mau, refaça a frase utilizando bem e bom. A forma equivalente a "bem" é "mal", e a forma equivalente a "bom" é "mau".

26. Suprima o "mais". O sentido de preferir é "querer mais". "Preferir mais" é redundante.

27. "Quem prefere, prefere alguma coisa a alguma coisa". Não se usa "do que". A preposição adequada é "a".

28. As palavras de sentido negativo atraem o pronome átono para antes do verbo.

29. Palavras negativas atraem o pronome átono para antes do verbo.

30. Os pronomes relativos e as conjunç oes subordinativas atraem o pronome para antes do verbo.

31. Certos advérbios (sempre, já, bem, aqui, onde, mais, talvez, ainda, como, por que) atraem o pronome para antes do verbo.

32. Os pronomes indefinidos "tudo, pouco, algo" atraem o pronome para antes do verbo.

33. "só, ou, ora e quer" atraem o pronome para antes do verbo.

34. Conjugação de um verbo irregular no futuro do subjuntivo.

35. O verbo "ir" constrói-se com preposição "a". Se o complemento (lugar a que se vai) for feminino, teremos crase.

36. O verbo "aderir" constrói-se com a preposição "a". Se o complemento (aquilo a que se adere) for feminino, teremos crase.

37. O verbo "pertencer" constrói-se com a preposição "a". Se o complemento (pertence a algo, ou a alguém) for feminino, teremos crase.

38. O verbo "candidatar-se" constrói-se com preposição "a". Se o complemento (aquilo a que o sujeito se candidata) for feminino, teremos crase.

39. O verbo obedecer/desobedecer constrói-se com a preposição a. Se o complemento (obedecer a quem) for feminino, teremos crase.

40. O adjetivo na função de predicativo concorda em gênero (masculino ou feminino) e número (singular ou plural) com o sujeito.

41. Quando um nome que rege preposição "a" é complementado por palavra feminina, ocorrerá crase.

42. O verbo "reagir" constrói-se com a preposição "a". Se o complemento (reage a alguma coisa) for feminino, teremos crase.

43. Os verbos "obedecer"/"desobedecer" constroem-se com preposição "a". Quem obedece, obedece a alguém.

44. Os pronomes pessoais "mim", "ti", "ele", "ela", "eles", "elas", "si", "nós", "nos", "vós", "vos" não admitem artigo, portanto não haverá crase antes deles.

45. O pronome "eu" não deve ser preposicionado. Use, nesse caso, "a mim".

46. Quem namora, namora alguém. Não use a preposição "com" na regência do verbo namorar.

47. Quando um nome que rege preposição "a" é complementado por palavra feminina plural determinada pelo artigo "as", ocorrerá crase.

48. Neste caso, o adjetivo "meia" concorda com o substantivo "hora", que está subentendido.

49. O que é equivalente, é equivalente a alguma coisa. Temos, portanto preposição a. Se o complemento for feminino, teremos crase.

50. O verbo "equivaler" constrói-se com a preposição "a". Se o complemento (equivale a algo) for feminino, teremos crase.

51. "Equivaler" constrói-se com prep. "a". Há crase com compl. feminino.

52. Os artigos definidos (o, a, os, as) e os artigos indefinidos (um, uma, uns, umas) concordam em número (singular ou plural) e em gênero (masculino ou feminino) com o substantivo a que se referem.

53. Os verbos "evitar" e "usufruir" não regem preposição "de". São transitivos diretos.

54. Atenção para a regência de alguns verbos: "demorar em" (em lugar de "demorar para"), "torcer por" (em lugar de "torcer para"), "votar em" (em lugar de "votar para").

55. O verbo "arrasar" não rege preposiçao "com". É transitivo direto.

56. Atençao para a regência do verbo "habituar-se": use "habituar-se a" em lugar de "habituar-se com".

57. A expressão correta é "‘a medida que".

58. O verbo "acarretar" é transitivo direto. É inadequado o uso da preposição "em".

59. Neste caso, o "a" é apenas preposição. Não se pode indicar crase.

60. O verbo "assistir" com o sentido de presenciar rege preposição "a". Se for seguido de palavra feminina singular, haverá crase.

61. "Valorização" rege preposição "de" e não preposição "a".

62. A expressão "ou seja" deve ser isolada por vírgulas.

63. A expressão "ou seja" deve ser isolada por vírgulas.

64. Deve haver vírgula antes de "no entanto", se esta expressar relação entre sentenças.

65. Os determinantes concordam em número (singular ou plural) e em gênero (masculino ou feminino) com o substantivo a que se referem.

66. Os numerais concordam em número (singular ou plural) e em gênero (masculino ou feminino) com o substantivo a que se referem.

67. Verificou-se erro de concordância entre o sujeito e o verbo.

68. Verifique o excesso de verbos em sequência.

69. O verbo "haver" já indica uma ocorrência no passado, portanto a palavra "atrás" é redundante.

70. "Conviver junto" ou "conviver juntos" são express oes redundantes. Suprima a palavra "junto" ou "juntos".

71. O adjetivo na função de predicativo concorda em gênero (masculino ou feminino) e número (singular ou plural) com o sujeito.

72. O sujeito concorda em número (singular ou plural) com o predicado.

73. Pronomes de tratamento não admitem artigo, portanto não haverá crase antes deles. A única exceção é o pronome "senhora", que admite artigo e, se houver também preposição, haverá crase.

74. O predicativo concorda em gênero (masculino ou feminino) e número (singular ou plural) com o sujeito.

75. O adjetivo na função de predicativo concorda em número (singular ou plural) com o verbo.

Appendix B

Conjunction for Subordinate Clauses

que, se, porque, já que, do que, uma vez que, embora, desde que, pois, a não ser que, enquanto que, se bem que, caso, dado que, tanto mais que, visto que, ainda que, como se, como que, e, mesmo que, uns vez que, ao passo que, de modo que, pois que, uma vez, de tal modo que, como, para que, sendo que, por, senão, por mais que, a fim de que

Appendix C

Conclusion Markers

portanto, logo, por conseguinte, pois, visto que, porquanto, então, por fim, afinal, finalmente, enfim, por isso, diante disso, é preciso, é necessário, é necessária, sendo necessário, sendo necessária, é importante, é essencial, faz-se necessária, dessa forma, desta forma, dessa maneira, desta maneira, deste jeito, uma forma, uma maneira, para se obter, para se alcançar, para se conseguir, para um melhor, para uma melhor, para que, para isso, para isto, para tanto, de fato, em suma, destarte, isto posto, à vista disso, em vista disso, assim sendo, sendo assim, consequentemente, dessarte, diante do exposto, a fim de, com intenç ao de, com o propósito de, com a finalidade de, com o intuito de, de modo a, de forma que, de maneira que, cabe

Appendix D

Shap Graphs

D.1 SHAP graphs

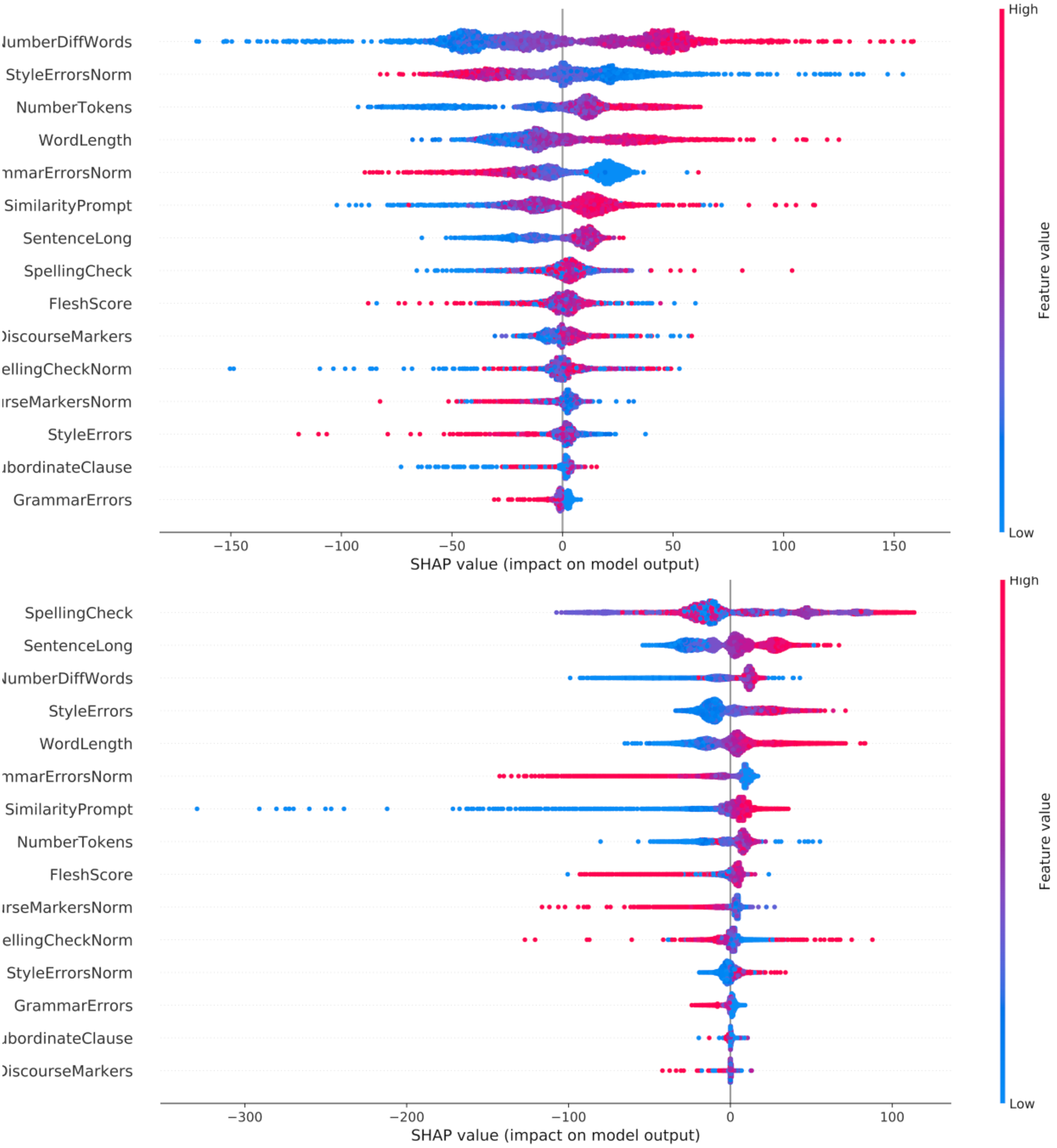


Figure D.1. SHAP graphs of the Baseline system showing the feature importance for the Final Grade of UOL Dataset (up) and Brasil Escola dataset (bottom)

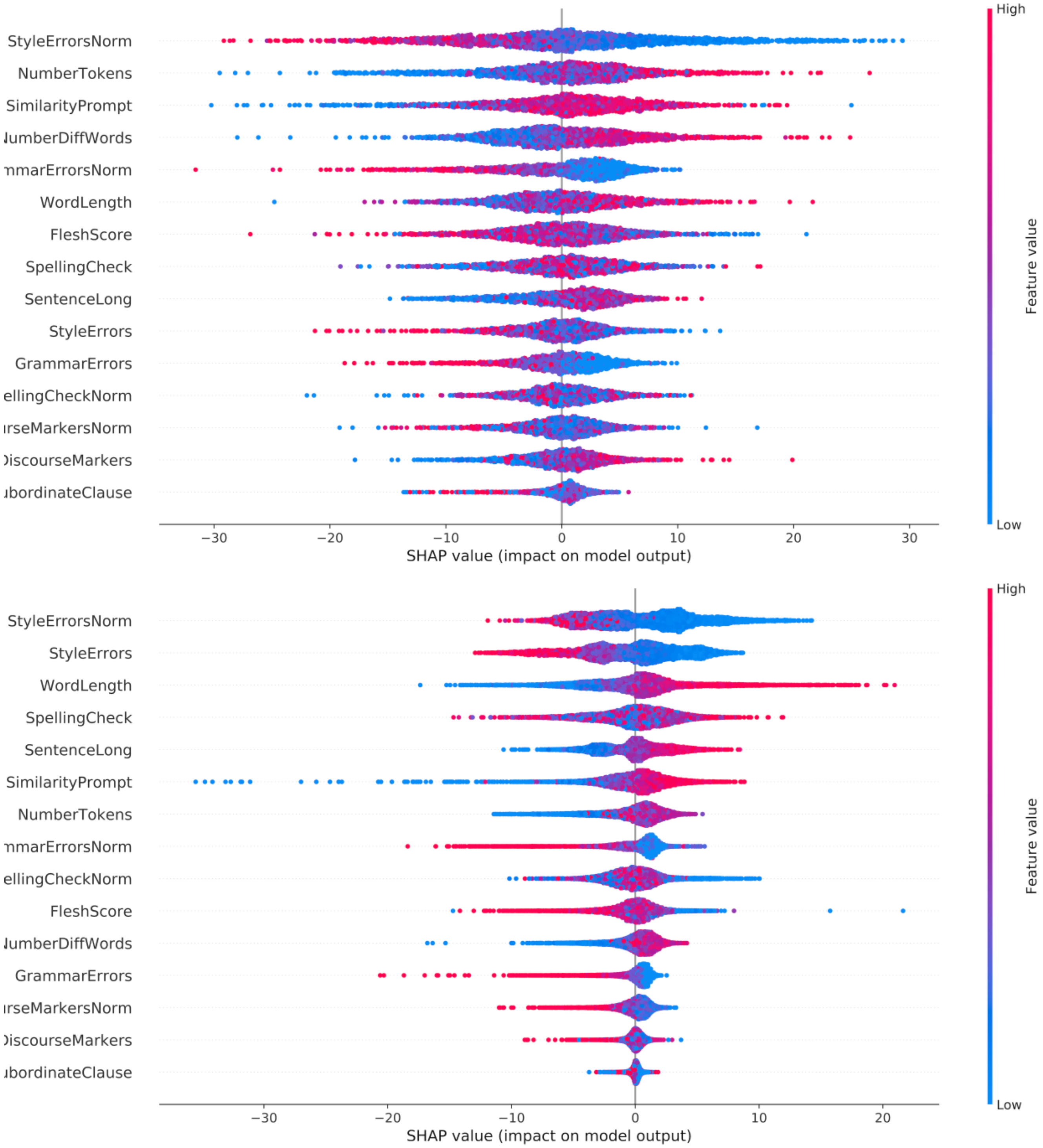


Figure D.2. SHAP graphs of the Baseline system showing the feature importance for the Formal Grade of UOL Dataset (up) and Brasil Escola dataset (bottom)

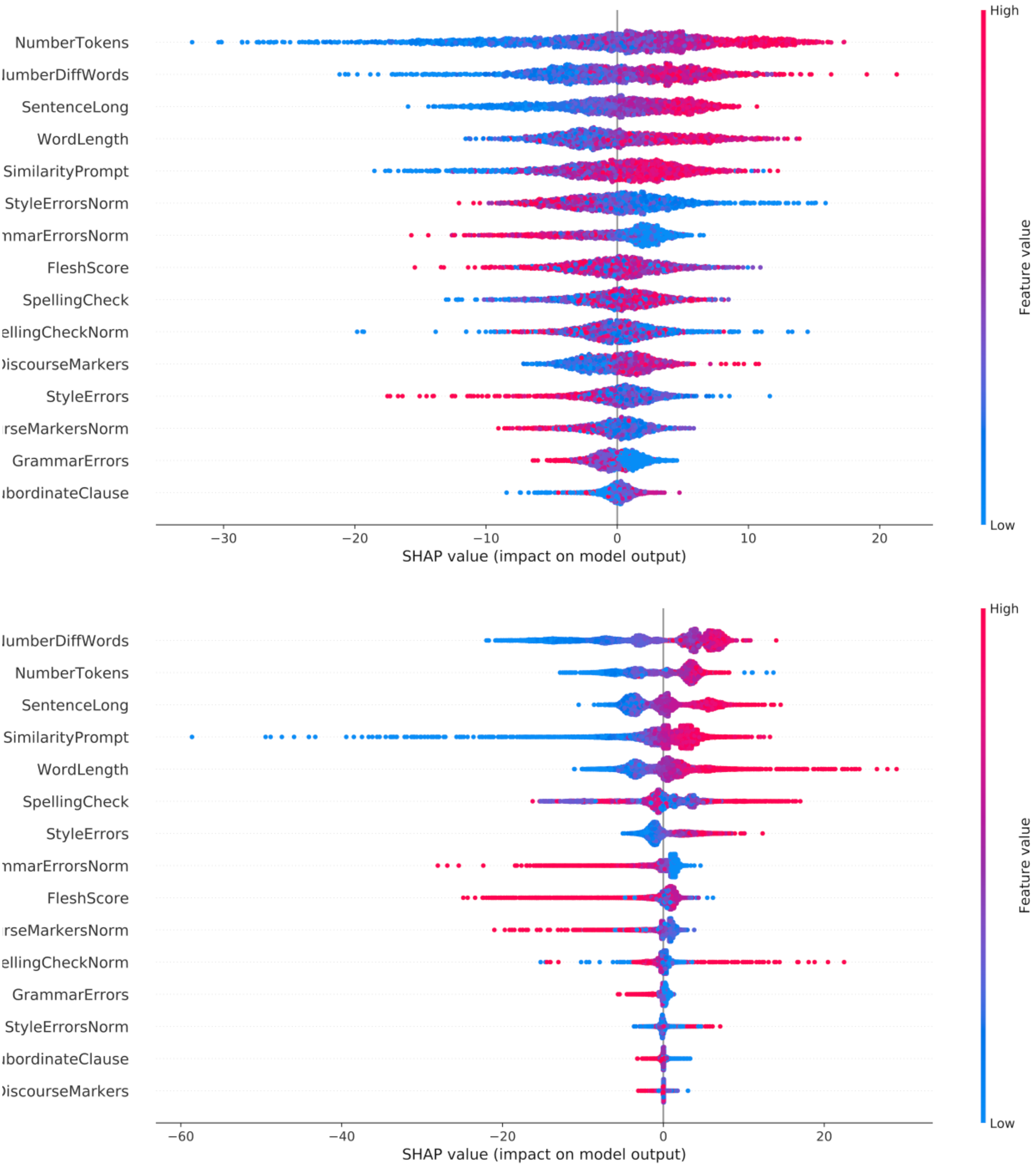
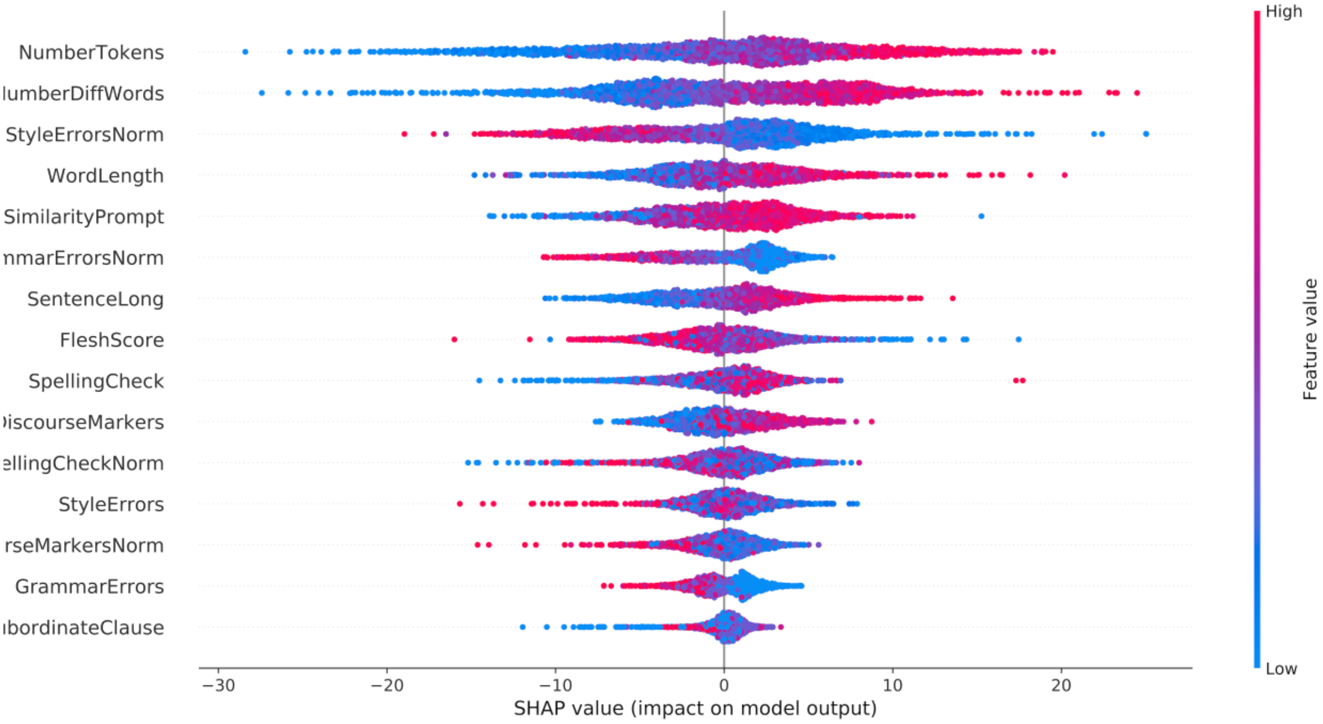


Figure D.3. SHAP graphs of the Baseline system showing the feature importance for the Understanding the Task Grade of UOL Dataset (up) and Brasil Escola dataset (bottom)



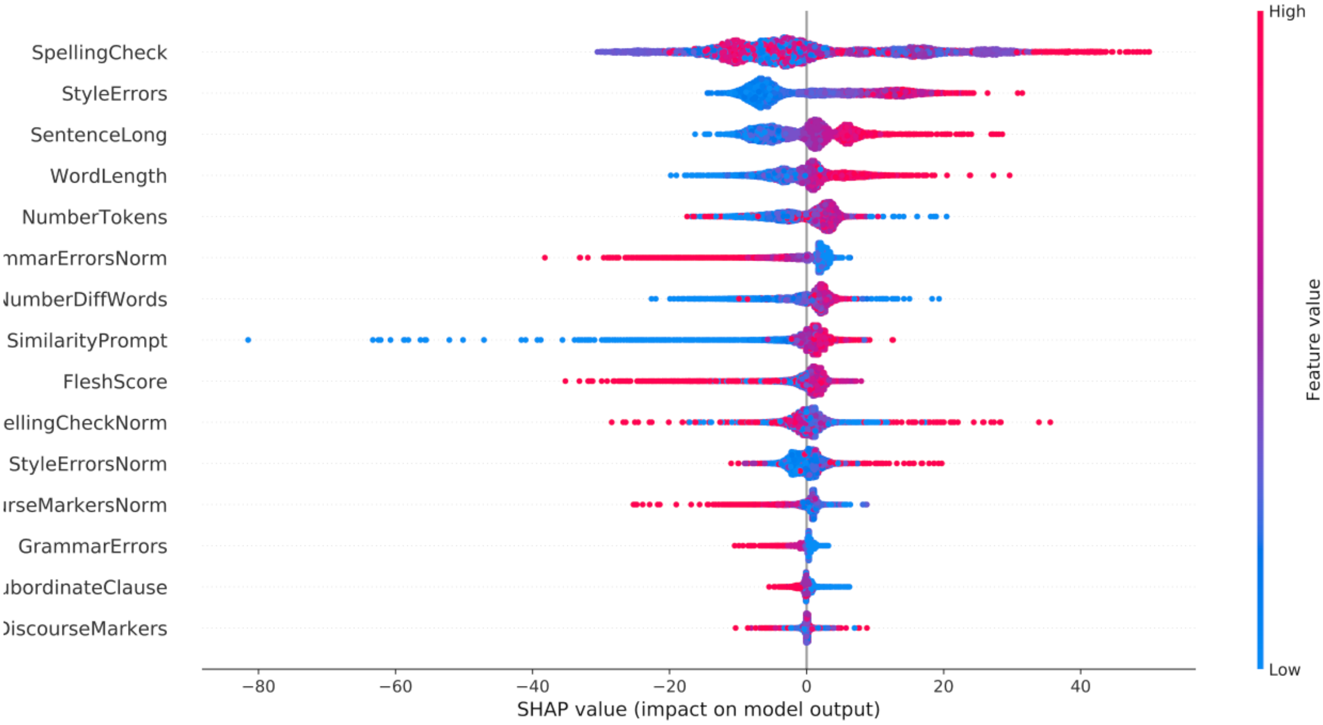
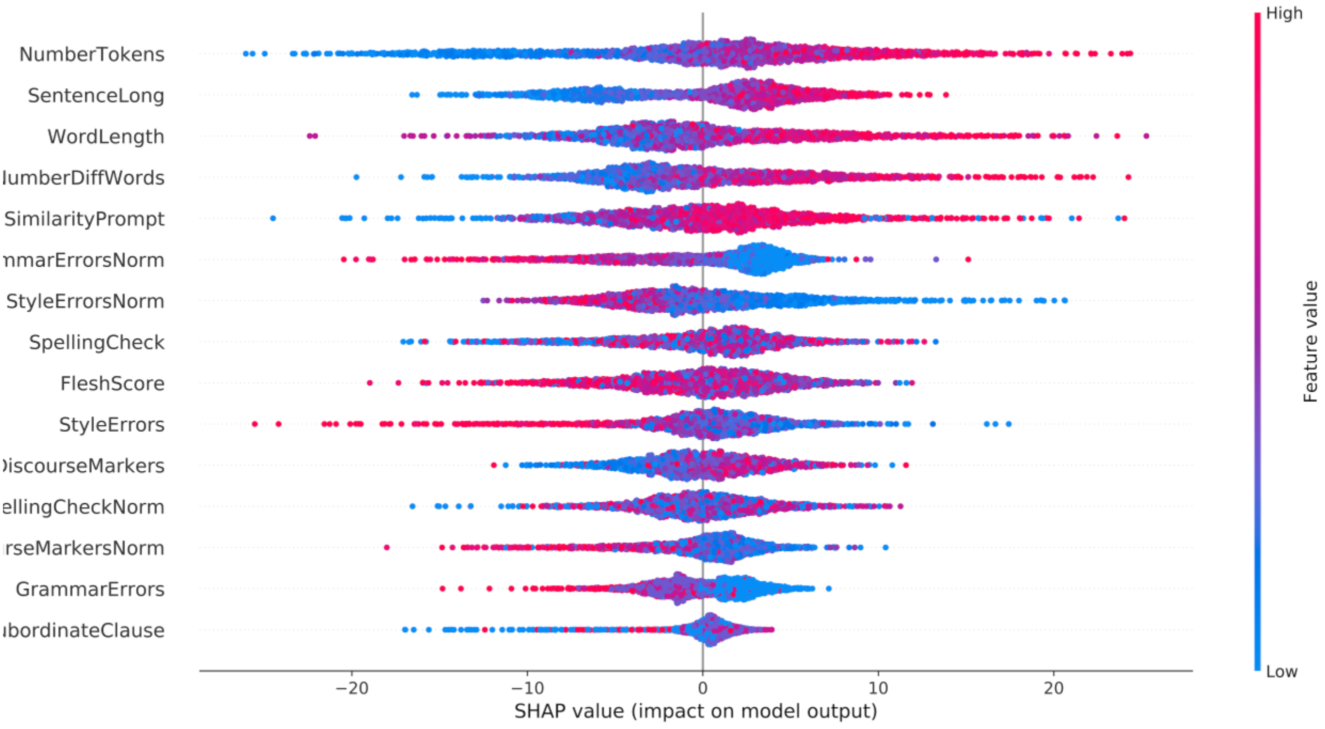


Figure D.4. SHAP graphs of the Baseline system showing the feature importance for the Organization of Information Grade of UOL Dataset (up) and Brasil Escola dataset (bottom)



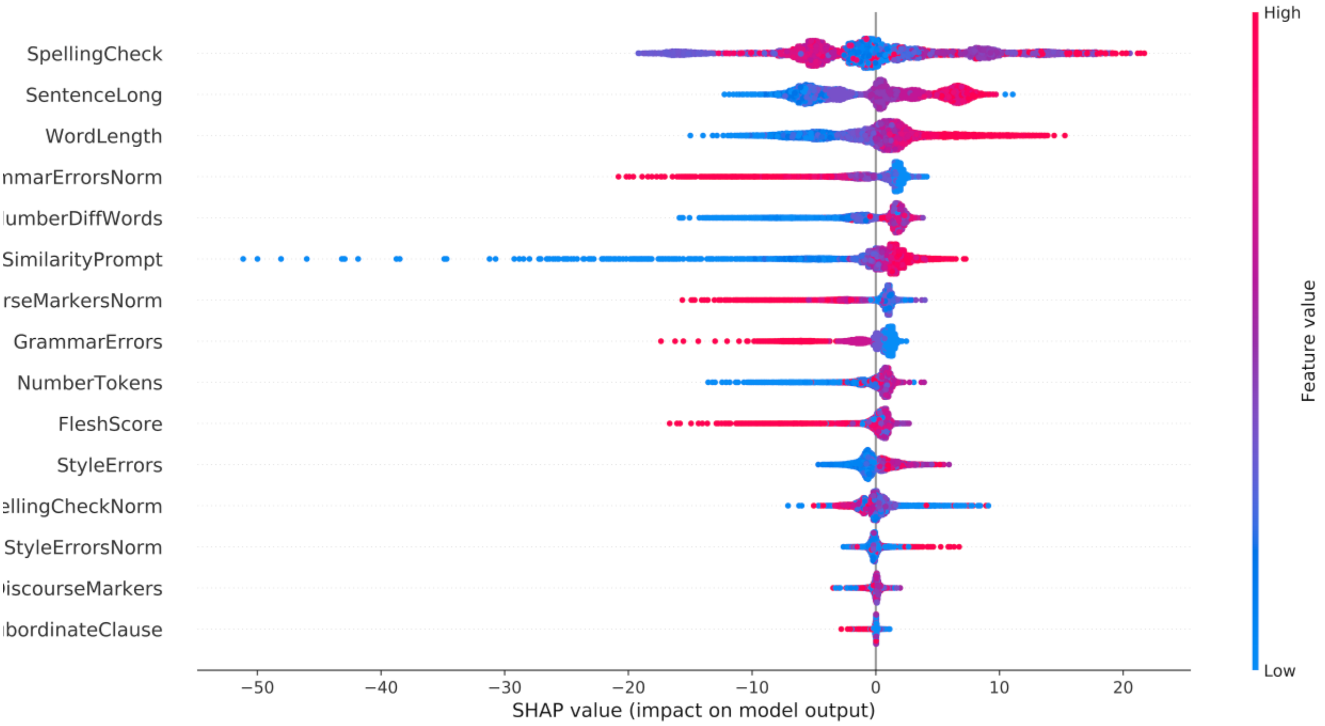
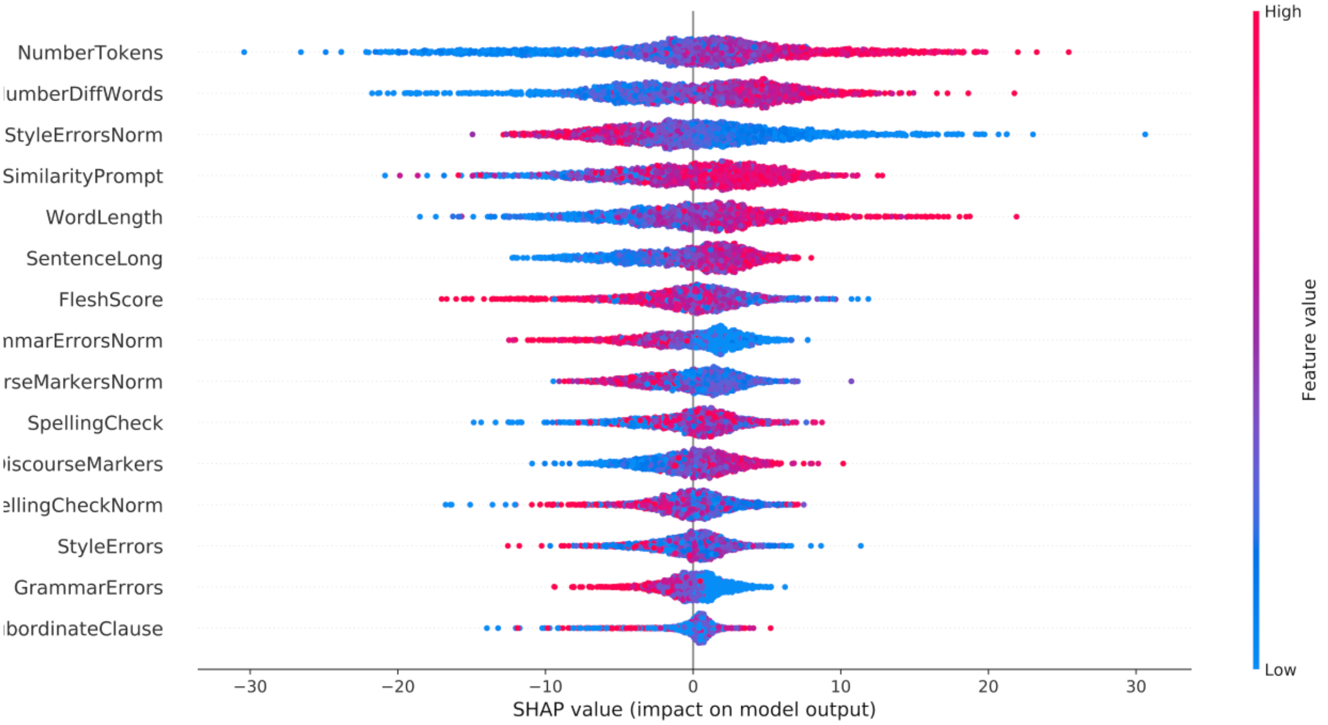


Figure D.5. SHAP graphs of the Baseline system showing the feature importance for the Knowing Argumentation grade of UOL Dataset (up) and Brasil Escola dataset (bottom)



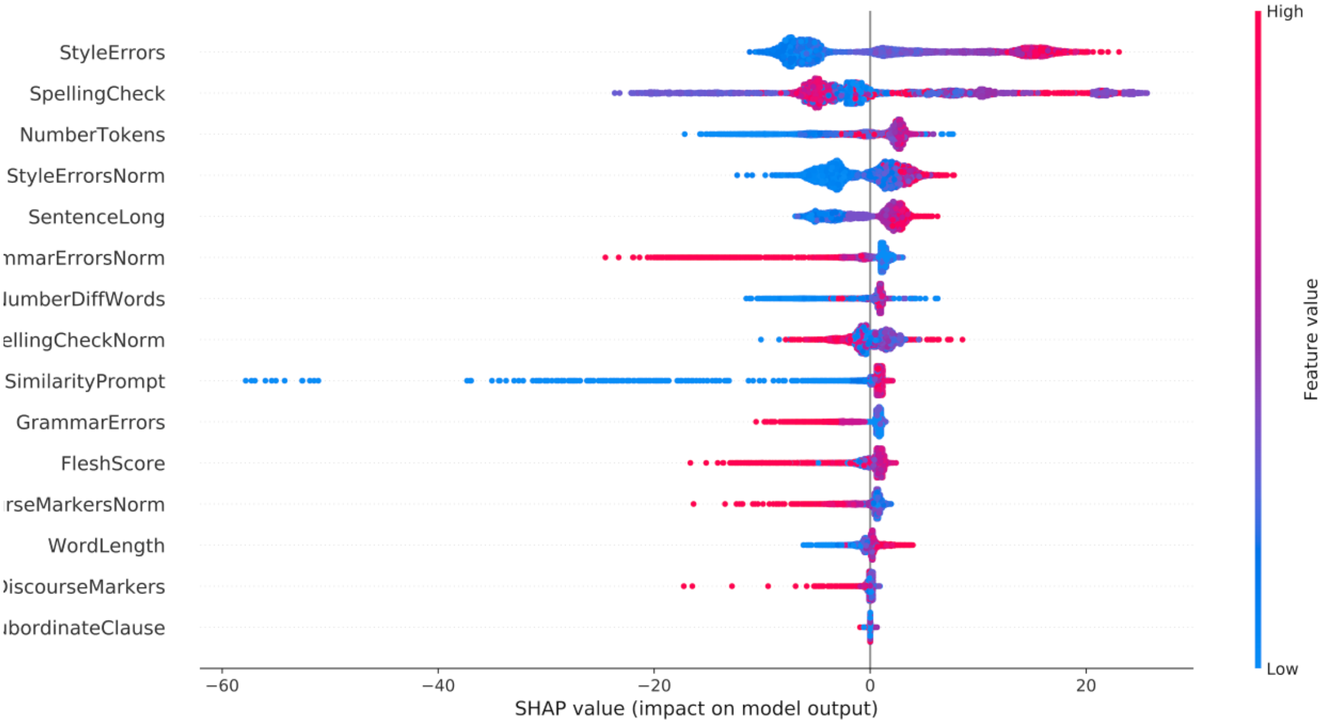


Figure D.6. SHAP graphs of the Baseline system showing the feature importance for the Solution Proposal grade of UOL Dataset (up) and Brasil Escola dataset (bottom)

Appendix E

Linear Models with Expanded Features

E.0.1 UOL

Aspect: Final Score

Features: FleshScore, FleshScore/NumberDiffWords, StyleErrorsNorm, NumberTokens/NumberDiffWords, StyleErrors/StyleErrorsNorm, NumberTokens/SubordinateClause, DiscourseMarkersNorm/SpellingCheckNorm, NumberTokens/SentenceLong, SpellingCheck/SpellingCheckNorm, Demonstrative/SimilarityPrompt, SentenceLong/NumberDiffWords, DiscourseMarkers/GrammarErrorsNorm, NumberDiffWords, StyleErrors/SimilarityPrompt, GrammarErrors/SentenceLong, WordLength, Enclise/NumberDiffWords, IndeternationInstrucments, Conjunctions, Demonstrative/GrammarErrors, GrammarErrors/RacistTerms, GrammarErrorsNorm/SubordinateClause, GrammarErrors/SpellingCheck, SpellingCheck/SimilarityPrompt, GrammarErrors/DiscourseMarkers, EncliseNorm, SpellingCheck/AnaphorPronouns, StyleErrors/AnaphorPronouns, FleshScore/EncliseNorm, SentenceLong/EncliseNorm, DiscourseMarkers/EncliseNorm, EncliseNorm/DiscourseMarkersNorm, Demonstrative/EncliseNorm

Aspect: Formal Language

Features: NumberDiffWords/AnaphorPronouns, StyleErrorsNorm/AnaphorPronouns, Demonstrative/SubordinateClause, FleshScore/NumberDiffWords, StyleErrorsNorm, Enclise/StyleErrorsNorm, Articles, DemonstrativeNorm, Enclise/StyleErrors, StyleErrors/StyleErrorsNorm, SpellingCheck/SubordinateClause, AnaphorPronouns, DiscourseMarkers/GrammarErrorsNorm, NumberDiffWords, StyleErrors/SimilarityPrompt, GrammarErrorsNorm/DiscourseMarkersNorm, SimilarityPrompt, FleshScore/SimilarityPrompt, WordLength, SentenceLong/AnaphorPronouns, GrammarErrors, IndeternationInstrucments, DiscourseMarkersNorm/SpellingCheckNorm, SentenceLong/EncliseNorm, FleshScore/EncliseNorm, Demonstrative/GrammarErrors, SentenceLong/GrammarErrorsNorm, GrammarErrors/RacistTerms, RacistTerms, NumberDiffWords/RacistTerms, SpellingCheckNorm/SubordinateClause, Demonstrative/StyleErrors

Aspect: Understanding the Task

Features: DemonstrativeNorm/StyleErrorsNorm, SimilaritySentencesMax, DemonstrativeNorm/DiscourseMarkersNorm, DemonstrativeNorm/AnaphorPronouns, SpellingCheck, SpellingCheck/SimilarityPrompt, SubordinateClause/RacistTerms, SpellingCheck/RacistTerms, SpellingCheck/AnaphorPronouns, EncliseNorm/StyleErrorsNorm, Enclise/SimilarityPrompt, DiscourseMarkers/EncliseNorm, SentenceLong/EncliseNorm, FleshScore/EncliseNorm, NumberTokens, DiscourseMarkersNorm/SpellingCheckNorm, SimilarityPrompt, NumberTokens/SimilarityPrompt, SimilarityPrompt/NumberDiffWords, WordLength, StyleErrors/SimilarityPrompt, IndeternationInstrucments, FleshScore/NumberDiffWords, GrammarErrors/EncliseNorm, GrammarErrors/SpellingCheck, SimilarityPrompt/StyleErrorsNorm, StyleErrors/NumberDiffWords, Demonstrative/DiscourseMarkersNorm, DiscourseMarkers/GrammarErrorsNorm, LexicalComplexity/AnaphorPronouns, SubordinateClause/AnaphorPronouns, Enclise/DiscourseMarkers, NumberDiffWords/DiscourseMarkersNorm, StyleErrors/AnaphorPronouns, SpellingCheckNorm, SpellingCheck/SubordinateClause, GrammarErrorsNorm/SubordinateClause, SentenceLong/GrammarErrorsNorm

Aspect: Organization of Information

Features: GrammarErrors/SentenceLong, WordLength, Enclise/NumberDiffWords, IndeternationInstrucments, NumberTokens/SubordinateClause, SentenceLong/GrammarErrorsNorm, DiscourseMarkersNorm/SpellingCheckNorm, SimilaritySentencesMax, Demonstrative/SimilarityPrompt, SimilarityPrompt/NumberDiffWords, SimilarityPrompt, NumberTokens/SimilarityPrompt, GrammarErrorsNorm/StyleErrorsNorm, GrammarErrors/RacistTerms, FleshScore/SubordinateClause, DemonstrativeNorm/GrammarErrorsNorm, StyleErrors, StyleErrors/NumberDiffWords, DiscourseMarkersNorm/AnaphorPronouns, StyleErrors/GrammarErrorsNorm, Enclise/DiscourseMarkersNorm, GrammarErrorsNorm/DiscourseMarkersNorm, SpellingCheckNorm, SimilarityPrompt/AnaphorPronouns

Aspect: Knowing Argumentation

Features: SimilarityPrompt/NumberDiffWords, DiscourseMarkersNorm, Enclise/DemonstrativeNorm, Demonstrative/SubordinateClause, SpellingCheckNorm, SpellingCheck, SpellingCheck/SimilarityPrompt, GrammarErrors/SentenceLong, Enclise, WordLength, IndeternationInstrucments, DiscourseMarkers/GrammarErrorsNorm, SentenceLong/SubordinateClause, StyleErrorsNorm, FleshScore/SimilarityPrompt, NumberTokens/NumberDiffWords, SpellingCheck/SpellingCheckNorm, NumberDiffWords/StyleErrorsNorm, StyleErrors, SimilarityPrompt, NumberDiffWords, GrammarErrors/RacistTerms, DemonstrativeNorm/LexicalComplexity, SimilaritySentencesMax

Aspect: Solution Proposal **Features:** StyleErrors, StyleErrors/NumberDiffWords, Enclise/SimilarityPrompt, FleshScore/StyleErrors, StyleErrorsNorm, FleshScore/SimilarityPrompt,, EncliseNorm/DiscourseMarkersNorm, Demonstrative/DiscourseMarkersNorm, SpellingCheck/NumberDiffWords, DemonstrativeNorm, Demonstrative/SentenceLong, ConclusionMarkers, SimilarityPrompt/StyleErrorsNorm, IndeternationInstrucments, NumberDiffWords/LexicalComplexity, SentenceLong/SubordinateClause, SimilarityPrompt/NumberDiffWords, SimilarityPrompt, NumberTokens/SimilarityPrompt, StyleErrorsNorm/RacistTerms, DiscourseMarkers/GrammarErrorsNorm, DiscourseMarkersNorm/SpellingCheckNorm, Pronouns, SpellingCheck/AnaphorPronouns, Enclise, WordLength, GrammarErrors/SubordinateClause

E.0.2 Brasil Escola

Aspect: Final Score

Features: StyleErrors/DiscourseMarkersNorm, NumberDiffWords/SpellingCheckNorm, DiscourseMarkers/SpellingCheck, NumberTokens/SpellingCheckNorm, StyleErrors/NumberDiffWords, SimilarityPrompt, Demonstrative/SentenceLong, SimilarityPrompt/NumberDiffWords, WordLength, SentenceLong, StyleErrors/StyleErrorsNorm, RacistTerms, SpellingCheck/EncliseNorm, IndeternationInstrucments, LexicalComplexity, NumberDiffWords, Conjunctions, DemonstrativeNorm/RacistTerms, GrammarErrorsNorm/StyleErrorsNorm, NumberTokens/StyleErrors, FleshScore/StyleErrorsNorm, Demonstrative/StyleErrorsNorm, DiscourseMarkersNorm/StyleErrorsNorm, DiscourseMarkers/DemonstrativeNorm, NumberTokens/SimilarityPrompt, DiscourseMarkers/LexicalComplexity, Demonstrative/DemonstrativeNorm, FleshScore/GrammarErrorsNorm, DemonstrativeNorm/DiscourseMarkersNorm, SentenceLong/DiscourseMarkersNorm, SentenceLong/DiscourseMarkers, DiscourseMarkersNorm, SentenceLong/GrammarErrorsNorm, GrammarErrorsNorm, Enclise/SentenceLong, StyleErrors/SubordinateClause, DiscourseMarkers/SubordinateClause, GrammarErrorsNorm/AnaphorPronouns, GrammarErrors/AnaphorPronouns, SentenceLong/NumberDiffWords, NumberTokens/StyleErrorsNorm, StyleErrorsNorm/RacistTerms, DiscourseMarkersNorm/RacistTerms, Enclise/RacistTerms, NumberTokens/LexicalComplexity, StyleErrors/GrammarErrorsNorm, Articles, Pronouns, FleshScore/StyleErrors

Aspect: Formal Language

Features: StyleErrors/SubordinateClause, RacistTerms, StyleErrors/RacistTerms, EncliseNorm/GrammarErrorsNorm, DiscourseMarkers/SpellingCheckNorm, SimilarityPrompt/SpellingCheckNorm, EncliseNorm/SpellingCheckNorm, DiscourseMarkersNorm/LexicalComplexity, SimilarityPrompt/LexicalComplexity, SentenceLong/DiscourseMarkersNorm, SentenceLong/DiscourseMarkers, SpellingCheck/LexicalComplexity, Enclise, EncliseNorm/RacistTerms, SentenceLong/EncliseNorm, SpellingCheckNorm/RacistTerms, DemonstrativeNorm/RacistTerms, SpellingCheck/GrammarErrorsNorm, NumberTokens/GrammarErrors, WordLength, SentenceLong, GrammarErrorsNorm/AnaphorPronouns, SentenceLong/AnaphorPronouns, EncliseNorm/AnaphorPronouns, DiscourseMarkers/EncliseNorm, EncliseNorm/SubordinateClause, SimilarityPrompt, StyleErrorsNorm, FleshScore/SimilarityPrompt, SimilarityPrompt/NumberDiffWords, StyleErrors, StyleErrors/GrammarErrorsNorm, Demonstrative, StyleErrors/DiscourseMarkersNorm, SimilarityPrompt/DemonstrativeNorm, NumberDiffWords/LexicalComplexity, NumberDiffWords, Demonstrative/DiscourseMarkers, FleshScore/GrammarErrorsNorm, GrammarErrorsNorm, Articles, FleshScore/StyleErrorsNorm, FleshScore/StyleErrors, SpellingCheck/EncliseNorm, NumberDiffWords/DemonstrativeNorm, GrammarErrors/DiscourseMarkersNorm, Adverbs, DiscourseMarkers/SimilarityPrompt, StyleErrors/SimilarityPrompt

Aspect: Understanding the Task

Features: SimilarityPrompt, WordLength, SentenceLong, StyleErrors/NumberDiffWords, StyleErrors/StyleErrorsNorm, Pronouns, LexicalComplexity, SimilarityPrompt/NumberDiffWords, EncliseNorm/SpellingCheckNorm, RacistTerms, DemonstrativeNorm/RacistTerms, NumberTokens/StyleErrors, Enclise/SentenceLong, IndeternationInstrucments, DiscourseMarkersNorm, GrammarErrorsNorm/StyleErrorsNorm, Demonstrative/DiscourseMarkersNorm, Demonstrative/DiscourseMarkers, Demonstrative/GrammarErrors, DemonstrativeNorm/GrammarErrorsNorm, GrammarErrors/NumberDiffWords, NumberDiffWords/StyleErrorsNorm, NumberTokens/SentenceLong, NumberTokens/StyleErrorsNorm, SentenceLong/DemonstrativeNorm, DiscourseMarkers/DemonstrativeNorm, SpellingCheck/GrammarErrorsNorm, SpellingCheckNorm/StyleErrorsNorm, SpellingCheck/StyleErrorsNorm, Enclise/DiscourseMarkersNorm, GrammarErrorsNorm/SubordinateClause, StyleErrorsNorm/SubordinateClause, EncliseNorm/SubordinateClause, DemonstrativeNorm/DiscourseMarkersNorm, DiscourseMarkers/StyleErrorsNorm, StyleErrors/DiscourseMarkersNorm, FleshScore, SentenceLong/RacistTerms, SimilarityPrompt/RacistTerms, Demonstrative/GrammarErrorsNorm, SentenceLong/NumberDiffWords

Aspect: Organization of Information

Features: Conjunctions, SpellingCheckNorm/RacistTerms, SpellingCheck/EncliseNorm, SimilarityPrompt/GrammarErrorsNorm, StyleErrorsNorm/RacistTerms, ConclusionMarkers, DiscourseMarkers/SpellingCheckNorm, NumberDiffWords/SpellingCheckNorm, GrammarErrorsNorm/SpellingCheckNorm, GrammarErrorsNorm/StyleErrorsNorm, GrammarErrors/StyleErrorsNorm, SentenceLong/DemonstrativeNorm, DiscourseMarkers/DemonstrativeNorm, DemonstrativeNorm/DiscourseMarkersNorm, SpellingCheck/DiscourseMarkersNorm, DiscourseMarkersNorm, FleshScore/RacistTerms, SentenceLong/GrammarErrorsNorm, FleshScore/GrammarErrorsNorm, DiscourseMarkers/NumberDiffWords, DiscourseMarkers/SubordinateClause, StyleErrors/SubordinateClause, NumberTokens/SimilarityPrompt, GrammarErrors/AnaphorPronouns, FleshScore/SubordinateClause, GrammarErrorsNorm/AnaphorPronouns, SentenceLong/DiscourseMarkersNorm, GrammarErrorsNorm, NumberTokens/SentenceLong, FleshScore, StyleErrors/RacistTerms, Demonstrative/SentenceLong, SimilarityPrompt, SentenceLong, WordLength, DiscourseMarkers/DiscourseMarkersNorm, SimilarityPrompt/NumberDiffWords, GrammarErrors/NumberDiffWords, StyleErrors/NumberDiffWords, Pronouns, LexicalComplexity, RacistTerms, IndeternationInstrucments, StyleErrors/DiscourseMarkersNorm, DemonstrativeNorm/AnaphorPronouns, SentenceLong/LexicalComplexity

Aspect: Knowing Argumentation

Features: SimilarityPrompt, StyleErrors/NumberDiffWords, DemonstrativeNorm/DiscourseMarkersNorm, StyleErrors/DiscourseMarkersNorm, StyleErrors/GrammarErrorsNorm, DiscourseMarkersNorm/LexicalComplexity, SimilarityPrompt/LexicalComplexity, DiscourseMarkers/DemonstrativeNorm, Pronouns, Adverbs, Articles, DiscourseMarkers/SimilarityPrompt, DiscourseMarkers/SubordinateClause, SimilarityPrompt/SubordinateClause, SimilarityPrompt/EncliseNorm, Enclise/SentenceLong, NumberTokens/SpellingCheck, NumberDiffWords/StyleErrorsNorm, Demonstrative/StyleErrorsNorm, FleshScore/StyleErrorsNorm, GrammarErrorsNorm/StyleErrorsNorm, GrammarErrors, RacistTerms, SpellingCheck/EncliseNorm, StyleErrors/StyleErrorsNorm, SimilarityPrompt/NumberDiffWords, NumberDiffWords, NumberTokens/DemonstrativeNorm, SimilarityPrompt/DemonstrativeNorm, Conjunctions, LexicalComplexity, Demonstrative/RacistTerms, IndeternationInstrucments, GrammarErrorsNorm/SpellingCheckNorm, EncliseNorm/GrammarErrorsNorm, Demonstrative/AnaphorPronouns, StyleErrors/SpellingCheckNorm, StyleErrors/SubordinateClause, Demonstrative/SpellingCheckNorm, SentenceLong, WordLength, GrammarErrorsNorm, GrammarErrorsNorm/AnaphorPronouns, Enclise/SpellingCheckNorm

Aspect: Solution Proposal

Features: NumberTokens/EncliseNorm, NumberTokens/SentenceLong, FleshScore/GrammarErrorsNorm, StyleErrors/NumberDiffWords, SimilarityPrompt/NumberDiffWords, NumberDiffWords, StyleErrorsNorm/RacistTerms, LexicalComplexity, GrammarErrorsNorm, SentenceLong, SentenceLong/GrammarErrorsNorm, StyleErrorsNorm, IndeternationInstrucments, StyleErrors/DemonstrativeNorm, AnaphorPronouns/RacistTerms, Demonstrative/StyleErrorsNorm, NumberTokens/StyleErrorsNorm, DiscourseMarkersNorm/LexicalComplexity, NumberTokens/LexicalComplexity, SentenceLong/LexicalComplexity, SpellingCheck/StyleErrors, SimilarityPrompt/GrammarErrorsNorm, SpellingCheckNorm/StyleErrorsNorm, SpellingCheck/StyleErrorsNorm, Enclise/RacistTerms, NumberTokens/Enclise, GrammarErrors/StyleErrorsNorm, SimilaritySentences, SpellingCheck/RacistTerms, Demonstrative, DemonstrativeNorm/DiscourseMarkersNorm, FleshScore/DiscourseMarkersNorm, Demonstrative/DiscourseMarkersNorm, StyleErrors/SpellingCheckNorm, GrammarErrorsNorm/SpellingCheckNorm, Demonstrative/SpellingCheckNorm, SimilarityPrompt, StyleErrors/StyleErrorsNorm, RacistTerms, SpellingCheck/EncliseNorm, Demonstrative/SentenceLong, Conjunctions, StyleErrors/DiscourseMarkersNorm, GrammarErrors/DiscourseMarkers, NumberTokens/AnaphorPronouns

Appendix F

Portuguese Lexicon List of Biased Terms (developed by Cançado et al. [2019])

Lexicon	Category	Lexicon	Category
bem	opinion/value	preferível	opinion/value
bom	opinion/value	preferivelmente	opinion/value
com franqueza	opinion/value	principal	opinion/value
com pesar	opinion/value	principalmente	opinion/value
com prazer	opinion/value	ruim	opinion/value
é uma pena	opinion/value	vale a pena	opinion/value
em especial	opinion/value	a rigor	opinion/value
especialmente	opinion/value	absolutamente	opinion/value
excelente	opinion/value	altamente	opinion/value
felizmente	opinion/value	amplamente	opinion/value
francamente	opinion/value	amplo	opinion/value
infelizmente	opinion/value	ampla	opinion/value
lamentavelmente	opinion/value	aproximadamente	opinion/value
mal	opinion/value	aproximado	opinion/value
melhor	opinion/value	aproximada	opinion/value
ótimo	opinion/value	bastante	opinion/value
particularmente	opinion/value	categoricamente	opinion/value
pesarosamente	opinion/value	cerca de	opinion/value
pior	opinion/value	como um todo	opinion/value
preferia	opinion/value	completamente	opinion/value

Lexicon	Category	Lexicon	Category
considerável	opinion/value	maioria	opinion/value
consideravelmente	opinion/value	mais	opinion/value
definitivamente	opinion/value	menos	opinion/value
demais	opinion/value	meramente	opinion/value
demasiadamente	opinion/value	minimamente	opinion/value
demasiado	opinion/value	minoria	opinion/value
demasiada	opinion/value	muitíssimo	opinion/value
elevado	opinion/value	muito	opinion/value
elevada	opinion/value	muita	opinion/value
enorme	opinion/value	pequeno	opinion/value
enormemente	opinion/value	pequena	opinion/value
escassamente	opinion/value	plenamente	opinion/value
escasso	opinion/value	pleno	opinion/value
escassa	opinion/value	plena	opinion/value
estritamente	opinion/value	pobre	opinion/value
estrito	opinion/value	pouco	opinion/value
estrita	opinion/value	pouca	opinion/value
exageradamente	opinion/value	pouquíssimo	opinion/value
excessivamente	opinion/value	praticamente	opinion/value
excessivo	opinion/value	precisamente	opinion/value
excessiva	opinion/value	quase	opinion/value
exclusivamente	opinion/value	razoável	opinion/value
exclusivo	opinion/value	razoavelmente	opinion/value
exclusiva	opinion/value	relativamente	opinion/value
expressamente	opinion/value	relativo	opinion/value
extremamente	opinion/value	relativa	opinion/value
extremo	opinion/value	rico	opinion/value
extrema	opinion/value	rica	opinion/value
grande	opinion/value	rigorosamente	opinion/value
grandemente	opinion/value	significativamente	opinion/value
imensamente	opinion/value	significativo	opinion/value
imenso	opinion/value	significativa	opinion/value
imensa	opinion/value	simples	opinion/value
incrivelmente	opinion/value	simplesmente	opinion/value
incrível	opinion/value	tanto	opinion/value
levemente	opinion/value	tanta	opinion/value

Lexicon	Category	Lexicon	Category
tão	opinion/value	decididamente	possibility/necessity
tipicamente	opinion/value	desnecessariamente	possibility/necessity
totalmente	opinion/value	dificilmente	possibility/necessity
tremenda	opinion/value	efetivamente	possibility/necessity
tremendamente	opinion/value	evidentemente	possibility/necessity
alguns	opinion/value	exatamente	possibility/necessity
algumas	opinion/value	facultativamente	possibility/necessity
às vezes	opinion/value	fundamentalmente	possibility/necessity
comum	opinion/value	indubitavelmente	possibility/necessity
comumente	opinion/value	inegavelmente	possibility/necessity
constante	opinion/value	justamente	possibility/necessity
constantemente	opinion/value	logicamente	possibility/necessity
de modo geral	opinion/value	na realidade	possibility/necessity
de vez em quando	opinion/value	na verdade	possibility/necessity
em geral	opinion/value	naturalmente	possibility/necessity
eventualmente	opinion/value	necessariamente	possibility/necessity
frequente	opinion/value	notadamente	possibility/necessity
frequentemente	opinion/value	obrigatoriamente	possibility/necessity
generalizada	opinion/value	obviamente	possibility/necessity
generalizado	opinion/value	possivelmente	possibility/necessity
geralmente	opinion/value	precisamente	possibility/necessity
normal	opinion/value	predominantemente	possibility/necessity
normalmente	opinion/value	presumivelmente	possibility/necessity
ocasional	opinion/value	provavelmente	possibility/necessity
ocasionalmente	opinion/value	realmente	possibility/necessity
raramente	opinion/value	seguramente	possibility/necessity
raro	opinion/value	sem dúvida	possibility/necessity
rara	opinion/value	talvez	possibility/necessity
sempre	opinion/value	virtualmente	possibility/necessity
usual	opinion/value	é aconselhável	possibility/necessity
usualmente	opinion/value	é certo	possibility/necessity
aparentemente	possibility/necessity	é claro	possibility/necessity
basicamente	possibility/necessity	é conveniente	possibility/necessity
certamente	possibility/necessity	é desnecessário	possibility/necessity
claramente	possibility/necessity	é desnecessária	possibility/necessity
com certeza	possibility/necessity	é devido	possibility/necessity
de fato	possibility/necessity	é difícil	possibility/necessity

Lexicon	Category
é duvidoso	possibility/necessity
é duvidosa	possibility/necessity
é efetivo	possibility/necessity
é efetiva	possibility/necessity
é evidente	possibility/necessity
é esperado	possibility/necessity
é facultativo	possibility/necessity
é fato	possibility/necessity
é fundamental	possibility/necessity
é importante	possibility/necessity
é impossível	possibility/necessity
é improvável	possibility/necessity
é indubitável	possibility/necessity
é inegável	possibility/necessity
é justo	possibility/necessity
é lógico	possibility/necessity
é natural	possibility/necessity
é necessário	possibility/necessity
é necessária	possibility/necessity
é obrigatório	possibility/necessity
é obrigatória	possibility/necessity
é óbvio	possibility/necessity
é possível	possibility/necessity
é preciso	possibility/necessity
é presumível	possibility/necessity
é provável	possibility/necessity
é real	possibility/necessity
é recomendável	possibility/necessity
é relevante	possibility/necessity
é seguro	possibility/necessity
deva	possibility/necessity
deve	possibility/necessity
deverá	possibility/necessity
deverão	possibility/necessity
deveria	possibility/necessity
deve-se	possibility/necessity
devia	possibility/necessity
impossibilidade	possibility/necessity
necessidade	possibility/necessity
necessita	possibility/necessity
necessitamos	possibility/necessity
necessitaria	possibility/necessity
necessitou-se	possibility/necessity
obrigatoriedade	possibility/necessity
pode	possibility/necessity
podemos	possibility/necessity
poderia	possibility/necessity
poderíamos	possibility/necessity
pode-se	possibility/necessity
podia	possibility/necessity
podiam	possibility/necessity
possa	possibility/necessity
possibilidade	possibility/necessity
possibilita	possibility/necessity
possibilitando	possibility/necessity
posso	possibility/necessity
precisa	possibility/necessity
precisamos	possibility/necessity
precisaria	possibility/necessity
pudesse	possibility/necessity
recomendamos	possibility/necessity
recomenda-se	possibility/necessity
recomendação	possibility/necessity
recomendou-se	possibility/necessity
tem que	possibility/necessity
temos que	possibility/necessity
tem-se que	possibility/necessity
tendo que	possibility/necessity
terão que	possibility/necessity
teria que	possibility/necessity
teriam que	possibility/necessity
tinha que	possibility/necessity
tivesse que	possibility/necessity
tem de	possibility/necessity

Lexicon	Category	Lexicon	Category
temos de	possibility/necessity	esperando	possibility/necessity
tem-se de	possibility/necessity	esperou-se	possibility/necessity
tendo de	possibility/necessity	fala-se	possibility/necessity
terão de	possibility/necessity	falou-se	possibility/necessity
teria de	possibility/necessity	falando-se	possibility/necessity
teriam de	possibility/necessity	ficaria	possibility/necessity
tinha de	possibility/necessity	gostaria	possibility/necessity
tivesse de	possibility/necessity	gostaríamos	possibility/necessity
basta	possibility/necessity	imaginamos	possibility/necessity
bastaria	possibility/necessity	imaginaríamos	possibility/necessity
bastasse	possibility/necessity	imagino	possibility/necessity
bastava	possibility/necessity	imaginando	possibility/necessity
acaba que	possibility/necessity	limita-se	possibility/necessity
acaba sendo	possibility/necessity	limitou-se	possibility/necessity
achamos	possibility/necessity	nada impede	possibility/necessity
achando	possibility/necessity	parece	possibility/necessity
acho	possibility/necessity	pareceu	possibility/necessity
acreditamos	possibility/necessity	parecendo	possibility/necessity
acredita-se	possibility/necessity	pensamos	possibility/necessity
acreditando	possibility/necessity	pensaríamos	possibility/necessity
acredito	possibility/necessity	pensando	possibility/necessity
acreditou-se	possibility/necessity	penso	possibility/necessity
certeza	possibility/necessity	procura-se	possibility/necessity
creio	possibility/necessity	procuramos	possibility/necessity
crendo	possibility/necessity	procurando	possibility/necessity
crença	possibility/necessity	procurou-se	possibility/necessity
dúvida	possibility/necessity	quer	possibility/necessity
duvida-se	possibility/necessity	queremos	possibility/necessity
duvido	possibility/necessity	querendo	possibility/necessity
duvidamos	possibility/necessity	queria	possibility/necessity
duvidando	possibility/necessity	quis	possibility/necessity
é dever	possibility/necessity	será	possibility/necessity
estou certo	possibility/necessity	seria	possibility/necessity
estamos certo	possibility/necessity	supomos	possibility/necessity
estando certo	possibility/necessity	suponho	possibility/necessity
espera-se	possibility/necessity	supondo	possibility/necessity
esperamos	possibility/necessity	adivinhamos	pressuposition

Lexicon	Category	Lexicon	Category
adinhando	pressuposition	lamentando	pressuposition
adivinha-se	pressuposition	lamenta-se	pressuposition
agora	pressuposition	lamentou-se	pressuposition
ainda	pressuposition	lastimamos	pressuposition
antes que	pressuposition	lastimando	pressuposition
atualmente	pressuposition	lastima-se	pressuposition
começa a	pressuposition	lastimou-se	pressuposition
começando a	pressuposition	lembramos	pressuposition
começou a	pressuposition	lembrando	pressuposition
como se sabe	pressuposition	lembre-se	pressuposition
constata-se	pressuposition	levando-se em conta	pressuposition
constatamos	pressuposition	no momento	pressuposition
constatando	pressuposition	notamos	pressuposition
constatou-se	pressuposition	notando	pressuposition
continua a	pressuposition	nota-se	pressuposition
continuando a	pressuposition	note-se	pressuposition
continuou a	pressuposition	notou-se	pressuposition
deixa de	pressuposition	observamos	pressuposition
deixou de	pressuposition	observa-se	pressuposition
deixando de	pressuposition	observando	pressuposition
depois que	pressuposition	observou-se	pressuposition
por mais que	pressuposition	parando	pressuposition
descobre-se	pressuposition	parou de	pressuposition
descobrimos	pressuposition	pelo que se sabe	pressuposition
descobriu-se	pressuposition	percebemos	pressuposition
desde que	pressuposition	percebe-se	pressuposition
diz respeito a	pressuposition	percebendo	pressuposition
é bom que	pressuposition	percebeu-se	pressuposition
é sabido	pressuposition	perdeu	pressuposition
hoje em dia	pressuposition	perdendo	pressuposition
ignora-se	pressuposition	posto que	pressuposition
ignoramos	pressuposition	reconhecemos	pressuposition
ignorando	pressuposition	reconhece-se	pressuposition
ignorou-se	pressuposition	reconhecendo	pressuposition
iniciou em	pressuposition	reconheceu-se	pressuposition
já	pressuposition	reparamos	pressuposition
lamentamos	pressuposition	repara-se	pressuposition

Lexicon	Category	Lexicon	Category
reparando	pressuposition	quando menos	argumentative
reparou-se	pressuposition	quando muito	argumentative
sabemos	pressuposition	sequer	argumentative
sabe-se	pressuposition	só	argumentative
sabendo	pressuposition	somente	argumentative
soube-se	pressuposition	a par disso	argumentative
sente-se	pressuposition	ademais	argumentative
sentimos	pressuposition	afinal	argumentative
sentindo	pressuposition	ainda	argumentative
sentiu-se	pressuposition	além	argumentative
trata-se	pressuposition	aliás	argumentative
tratou-se	pressuposition	como	argumentative
tratando-se	pressuposition	e	argumentative
vale lembrar	pressuposition	e não	argumentative
vale notar	pressuposition	em suma	argumentative
vale observar	pressuposition	enfim	argumentative
veja que	pressuposition	mas também	argumentative
vemos	pressuposition	muito menos	argumentative
vendo	pressuposition	não só	argumentative
vê-se	pressuposition	nem	argumentative
vimos	pressuposition	ou mesmo	argumentative
visto que	pressuposition	por sinal	argumentative
viu-se	pressuposition	também	argumentative
a ponto	argumentative	tampouco	argumentative
ao menos	argumentative	assim	argumentative
apenas	argumentative	com isso	argumentative
até	argumentative	como consequência	argumentative
até mesmo	argumentative	consequentemente	argumentative
incluindo	argumentative	de modo que	argumentative
inclusive	argumentative	deste modo	argumentative
mesmo	argumentative	em decorrência	argumentative
não mais que	argumentative	então	argumentative
nem mesmo	argumentative	logicamente	argumentative
no mínimo	argumentative	logo	argumentative
o único	argumentative	nesse sentido	argumentative
a única	argumentative	pois	argumentative
pelo menos	argumentative	por causa	argumentative

Lexicon	Category
por conseguinte	argumentative
por essa razão	argumentative
por isso	argumentative
portanto	argumentative
sendo assim	argumentative
ou	argumentative
ou então	argumentative
ou mesmo	argumentative
nem	argumentative
como se	argumentative
de um lado	argumentative
por outro lado	argumentative
mais que	argumentative
menos que	argumentative
nada mais que	argumentative
não só	argumentative
tanto	argumentative
quanto	argumentative
tão	argumentative
como	argumentative
desde que	argumentative
do contrário	argumentative
em lugar	argumentative
em vez	argumentative
enquanto	argumentative
no caso	argumentative
quando	argumentative
se	argumentative
se acaso	argumentative
senão	argumentative
de certa forma	argumentative
desse modo	argumentative
em função	argumentative
enquanto	argumentative
isso é	argumentative
já que	argumentative
na medida que	argumentative

Lexicon	Category
nessa direção	argumentative
no intuito	argumentative
no mesmo sentido	argumentative
ou seja	argumentative
pois	argumentative
porque	argumentative
que	argumentative
uma vez que	argumentative
tanto que	argumentative
visto que	argumentative
ainda que	argumentative
ao contrário	argumentative
apesar de	argumentative
contrariamente	argumentative
contudo	argumentative
embora	argumentative
entretanto	argumentative
fora isso	argumentative
mas	argumentative
mesmo que	argumentative
não obstante	argumentative
não fosse isso	argumentative
nem por isso	argumentative
no entanto	argumentative
para tanto	argumentative
pelo contrário	argumentative
por sua vez	argumentative
porém	argumentative
posto que	argumentative
todavia	argumentative