An Approach to Assess the Existence of a Proposed Intervention in Essay-argumentative Texts

Jonathan Nau[1], Aluizio Haendchen Filho[1], Fernando Concatto[1], Hercules Antonio do Prado[2], Edilson Ferneda[2] and Rudimar Luis Scaranto Dazzi[1]
[1]*Laboratory of Applied Intelligence, University of the Itajai Valley (UNIVALI), Rua Uruguay, 458, Itajai, Brazil*
[2]*MGTI, Catholic University of Brasilia (UCB), QS 07 - Lote 01, EPCT, Bl. K, sala 248, Taguatinga, Brazil*

Keywords: Automatic Essays Scoring, Argumentation Mining, ENEM, Machine Learning.

Abstract: This paper presents an approach for grading essays based on the presence of one or more theses, arguments, and intervention proposals. The research was developed by means of the following steps: *(i)* corpus delimitation and annotation; *(ii)* features selection; *(iii)* extraction of the training corpus, and *(iv)* class balancing, training and testing. Our study shows that features related to argumentation mining can improve the automatic essay scoring performance compared to the set of usual features. The main contribution of this paper is to demonstrate that argument marking procedures to improve score prediction in essays classification can produce better results. Moreover, it remained clear that essays classification does not depends on the number of features but rather on the ability of creating meaningful features for a given domain.

1 INTRODUCTION

One important goal of Artificial Intelligence (AI) is to make computers *act* just like humans, analyzing, reasoning, understanding, and proposing answers to different situations. Technological development comes to the simulation of human thoughts and actions, thanks to a specific AI related field known as Cognitive Computing. Advances in Cognitive Computing are making possible to attribute to machines the ability to infer the semantic relationship for any kind of data - text, audio or video. One can expect that, over time, the machine can *understand* the meaning of any document, extracting the value of information, relating it and assisting in decision making (Florão, 2017). Cognitive computing is the simulation of the human thinking process in a computerized way. It involves computerized platforms for machine learning, pattern recognition, case-based reasoning, natural language processing, computational linguistics, among other technologies.

Research in Linguistics and Computational Linguistics has long proven that a discourse is more than just a sequence of juxtaposed sentences. It comprises a linguistic production of more than one sentence, and its understanding is performed by considering the text to be understood as a whole. A text has a highly elaborated underlying structure that relates all its content, giving it coherence. This structure is called discursive structure, which is the object of study of the research area known as Discourse Analysis.

Understanding speech represents an essential component for students during the learning process. Linking information consistently, while organically building a solid knowledge base, is crucial for student development, but requires regular assessment and monitoring of progress.

The argumentative practice enables the student to articulate knowledge in order to develop a consistent reasoning in defending a point of view, thus mobilizing the basis for understanding a phenomenon. By presenting the argument, the student brings to the teacher and the class group their understanding of the phenomenon. Thus, the teacher has elements to evaluate this understanding and have subsidies to continue the pedagogical work. We realize that the written texts are still marked by orality. In summary, we find that the characteristics of scientific literacy are found more frequently in students' written argumentative productions than in those that do not assume such a

Nau, J., Filho, A., Concatto, F., Antonio do Prado, H., Ferneda, E. and Dazzi, R.
An Approach to Assess the Existence of a Proposed Intervention in Essay-argumentative Texts.
DOI: 10.5220/0009795206660673
In *Proceedings of the 22nd International Conference on Enterprise Information Systems (ICEIS 2020) - Volume 1*, pages 666-673
ISBN: 978-989-758-423-7

configuration (Lira, 2009).

The use of discursive knowledge is a relevant issue for Natural Language Processing systems (NLP). According to Dias da Silva (1996), NLP is a complex and multifaceted domain which objective is the design and implementation of computer systems that perform actions like spell and grammar checking, writing aid, text summarization, automatic writing correction and to structure dialoguing systems.

The aim of this paper is to present an approach for automatic detection of argumentative structure in text productions. Automatic detection of essay arguments can be very valuable for teachers, students, and applications (Stede, Schneider, 2018). When incorporated into correction or auto-detection algorithms in essays, it can help teachers to improve the correction process. In addition, it enables argumentative text evaluation procedures to be applied on a larger scale, providing guidance for pedagogical work in many disciplines.

2 BACKGROUND

2.1 Argument Structure

There is no a unique definition of an argument reported in the literature related to Theory of Argumentation. According to Walton (2009), the minimum definition says that an argument is a set of propositions composed by three parts: *(i)* a conclusion; *(ii)* a set of assumptions; and *(iii)* an inference from the assumptions for the conclusion. In addition, a given argument may be supported or refuted by other arguments. According to Stab and Gurevych (2017), an argument consists of several argument components, including a claim and one or more premises. The claim (also called conclusion) is a controversial statement and the central component of an argument.

Johnson and Blair (1994) proposed three binary criteria, known as RAS-criteria, that a logically good argument needs to fulfil: *(i)* relevance: if all of its premises count in favor of the truth (or falsity) of the claim; *(ii)* acceptability: if its premises represent undisputed common knowledge or facts; *(iii)* sufficiency: if its premises provide enough evidence for accepting or rejecting the claim.

In the Brazilian National High School Examination (ENEM) the production of a dissertative-argumentative text about a previously informed theme of social, scientific, cultural or political nature is required. This is the main test applied in Brazil to evaluate the writing skills of high school students.

In writing the essay, it is necessary: *(i)* to present a thesis - an opinion about the previously proposed theme, which in turn must be supported by arguments and *(ii)* to elaborate a proposal for social intervention for the problem presented, respecting human rights (Brasil, 2018). Figure 1 illustrates the writing process in the ENEM model.

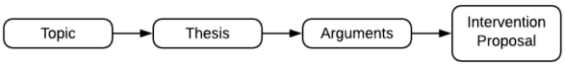

Figure 1: Process of preparing an essay on ENEM model.

Argument diagramming is one of the most important tools currently in use to assist with argument analysis and evaluation tasks. An argument diagram is essentially a graph representation of an argument in where the nodes contain propositions and the arrows are drawn from nodes to other nodes, representing inferences. Figure 2 presents the diagramming of arguments in an essay for Brazilian Portuguese according to the ENEM model. It shows the three components of an argument: the author's thesis, which will be defended during the writing, the arguments that will support the thesis, and finally, the intervention proposal.

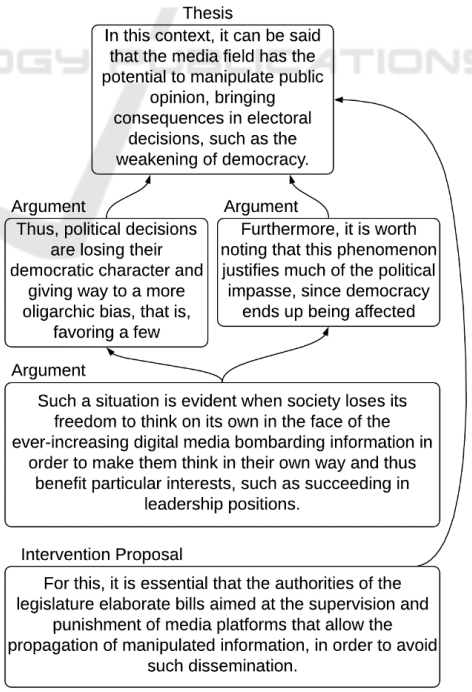

Figure 2: An Argument Diagram in the ENEM Model.

According to Stede and Schneider (2018), the classification of an Argumentative Discourse Unit (ADU) can occur in parts of a text or in the complete text as a single argumentative unit. In the case of all text as a single unit, the class is determined by its genre, such as newspapers, articles and essays. In addition, the identification of argumentative units can be applied to paragraphs or other passages of text; in general, the task is usually done at sentence or clause level. Therefore, it is possible to label each sentence or excerpt of the text, not only being limited to discovering its genre.

2.2 Argumentation Mining

Argumentation mining is the application of Natural Language Processing (NLP) tailored to identify and extract argumentative structures from texts (Stede, Schneider, 2018). For essays, the elements extracted are the thesis, the arguments, and the intervention proposal. In a scientific paper, for example, it is necessary to extract objectives, justifications, related works and results. Therefore, argumentation mining is not a standard process and depends strongly on the argumentative structure of a given text.

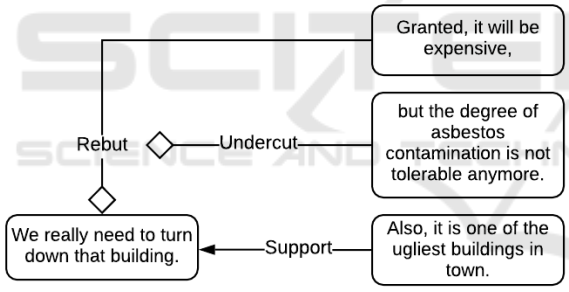

Figure 3: Example of an argumentative structure.

As shown in Figure 3, the fragment *We really need to tear down this building* indicates a statement by the author on a given subject; the proposition *will be expensive* presents an objection, but there is a counterargument in the fragment *but the degree contamination is no longer tolerable*. In the proposition *In addition, it is one of the ugliest buildings in the city* support is added to the main statement.

2.3 Inference Algorithms

Given a set of *n* real-valued *k*-dimensional vectors (independent variables), each associated with an expected value (dependent variable), one may employ supervised learning techniques to construct a model capable of predicting the output value of a previously unseen input vector. When the dependent variable is a real number, regression models are used; otherwise, when the dependent variable is an integer (which might represent a code for an object or concept), models based on classification are generally applied. In the context of essay, grades can be interpreted either as classes (for instance, good, bad and average) or as real numbers. In this work, the methods Gradient Boosted Trees (Friedman, 2001) and Support Vector Classification (SVC) (Knerr, Personnaz, Dreyfus, 1990) were applied. The choice for these methods were based on the results of Haendchen Filho et al (2019).

2.4 Imbalanced Learning

The problem of class imbalance occurs in many application domains, as in the case of essays. The imbalance of the number of samples among the classes presents a problem for traditional classification algorithms. The problem is that the classification algorithms try to maximize the accuracy of the classification without considering the meaning of the different classes. For example, if 25% of all essays are different from grade 1, then the algorithm will have 75% accuracy when classifying all essays as grade 1 (Seiffert et al, 2008).

After performing tests with different balancing techniques such as Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al, 2011), ADASYN (He et al, 2008), Random Oversampling (ROS) and Random Undersampling (RUS) (Yap et al, 2014), we chose to use the technique RUS, which presented the best results.

RUS approach aims to randomly discard data from the majority class to match the number of minority class examples. Undersampling is used when the amount of collected data is sufficient. Common methods of undersampling include cluster centroids and Tomek links (Batista et al, 2004), both of which target potential overlapping characteristics within the collected data sets to reduce the amount of majority data. The main disadvantage of this type of approach is that the probability of discarding useful data is very high.

3 RELATED WORKS

We have reviewed argument mining applied to text evaluation systems, especially the approaches that use end-to-end systems. In this sense, the research papers from Nguyen, Litman (2018) and Ghosh et al (2016) deserve to be highlighted.

Nguyen and Litman (2018) aims to show how these systems can add value to the evaluation of newsrooms. The work based on the argument mining approach implements a pipeline paradigm. The authors improved the argument component identification (ACI) model from Stab and Gurevych (2017) with features derived from Nguyen and Litman (2015). For argument component classification (ACC) and support relationship identification, we implemented our models based on Nguyen and Litman (2016).

To validate the approach, they used two corpora for training and testing. The ASAP data corpus, that contains 3,589 separate essays in two themes, and the TOEFL11 corpus, that contains 8,097 distributed in eight distinct themes. For TOEFL11 data, 10-fold cross-validation was used and for the ASAP set we performed 5-fold cross-validation. The models were trained with the Weka logistic regression algorithm. The algorithm was fed with 33 features extracted by means of argument mining, including *total number of words in argument components*, *number of paragraphs containing argument components*, *number of paragraphs that have supporting relationships*, among others.

The corpora ASAP 1, ASAP 2, and TOEFL11 presented respectively 0.830, 0.689 and 0.611 of F-score (the closer to 1 the better the result).

The second work, described by Ghosh et al (2016), took as hypothesis if the argumentative structure of the essays can help predict the grades of each essay. The results from Stab and Gurevych (2014) were used as a reference to implement argument mining. Argument mining was applied for extracting features for predicting essay grades; a total of 14 features were extracted (e.g., *number of relationships and argumentative components*, *depth of the argument tree*).

The experiment considered 107 essays from TOEFL11 corpus and used 10-fold cross-validation. It was applied logistic regression for training, reaching a quadratic-weighted kappa coefficient of 0.737 (the closer to 1 the better).

4 METHODOLOGY

The research work was developed by means of the following steps: *(i)* Delimitation and Annotation of the Marking Corpus; *(ii)* features selection; *(iii)* Extraction of the Training Corpus and *(iv)* Class Balancing, Training and Testing. These steps are described as follows.

4.1 Features Selection

A set of 623 features (Table 1) were considered for building the models.

Table 1: The several dimensions of the features.

Type	Description
Lexicon diversity and statistical (84 metrics)	Metrics that indicate how varied is the use of the lexicon in textual production. They were calculated from the token-type ratio and encompassed content words, functional words, verbs, adjectives, pronouns, paragraph size, paragraphs per sentence, and so on.
Bag of words (70 metrics)	Bag of words based on an analogical dictionary, searching for categories of words that convey ideas such as cause-effect relations, formation of ideas, comparison of ideas, hypothesis, cause, circumstance, purpose, conjunctions of condition, consequence, explanation, among others.
Textual coherence (179 metrics)	Coherence is achieved through syntactical features such as the use of deictic, anaphoric and cataphoric elements or a logical tense structure. Among the features, we can cite as an example: average similarities between the sentences of the first paragraph, number of justification markers in the first paragraph, number of antithesis markers in the first paragraph, number of markers single antitheses in the first paragraph, number of conclusion markers in the last paragraph, and so on.
Textual cohesion (187 metrics)	In order to identify the referential cohesion relations in the text, several overlapping indexes were calculated. For example, overlapping names and pronouns between adjacent sentences and paragraphs, overlapping of adjectives, verbs, adverbs, words of content, among others.
Adherence to the theme (98 metrics)	The adequacy or pertinence to the theme refers to how much the content of an essay is related to the thematic proposal to which the essay was submitted. An essay with good adaptation to the theme consistently maintains the theme introduced in the thematic proposal and is free of irrelevant disagreements.
Argument structure (5 metrics)	**Number of theses, argument, intervention proposals, nonarguments, components (theses + arguments + intervention proposals**

The usual set of 618 metrics – five first lines in the table - was enriched with five new features related to argumentative structure. These features were defined on an experiment with a corpus of 50 essays (extract at random) from the Brazil Escola portal. This portal contains a writing base from ENEM, in which students are encouraged to submit essays on a particular topic, receiving feedbacks from specialists. Subsequently, the essays were annotated by two independent specialists. In the annotation process, we used the BRAT tool (Stenetorp et al, 2012), a specific web tool for text annotation.

The annotation task involved the steps: *(i)* definition of the argument elements and annotation specifications; *(ii)* annotation of the components; *(iii)* analysis of component inconsistencies; *(iv)* annotation of argumentative relations; *(v)* analysis of the inconsistencies of relations; and *(vi)* preparation of the final corpus.

The annotation of components and relationships were proceeded at different moments because the annotators were independent. At first, the argumentative components were identified. These components were then analyzed to detect divergences. Next, a corpus with the argumentative units was compiled. Finally, the relationships were defined and the divergences verified, thus resulting in the final version of the annotated corpus of essays in Portuguese. The correlation between the annotators in each argument component (Thesis, Argument, and Intervention Proposal), measured by Krippendorff's α coefficient (Krippendorff, 2004), are, respectively, 0.87, 0.91 and 0.95.

The annotation of the components had the overall average correlation of 0.92. Among the annotated components, it is possible to observe the high agreement between the annotators in the argument components and intervention proposal, but the thesis presented the lowest agreement among the evaluators. Table 2 presents the statistics for the final corpus.

Table 2: Final corpus statistics.

	Total	**Average/essay**
Sentences	659	13.18
Words (tokens)	20,659	413.18
Thesis	62	1.24
Arguments	222	4.44
Intervention proposals	100	2
Not argumentative	275	5.5

Similarly to Almeida Júnior, Spalenza and Oliveira (2017), each essay was represented as a feature vector with the 623 features of Table 1.

The standardization of the statistical distribution of features directly influences the quality of the machine learning models because it reduces the negative effect that outliers may cause during the training process. So, to ensure the good performance of the model, *z*-score standardization was applied.

4.2 Extraction of the Training Corpus

The essays used to compose the corpus were obtained by means of a crawling process of essays datasets from the UOL 5 and Brazil School portal. Both portals have similar processes for persistence of essays: monthly, a theme is proposed and interested students submit their textual productions for evaluation. Part of the essays evaluated are then made available on the portal along with the respective corrections, scores and comments of the reviewers. For each essay, a score between 0 and 2 is assigned, varying in steps of 0.5 for the five competences corresponding to the ENEM evaluation model. As we can see next, this scoring approach is very similar to the ENEM official scoring system.

The five competencies evaluated are: *(i)* demonstration of mastering on the formal written of Portuguese Language; *(ii)* understanding the essay proposal within the structural limits of the essay-argumentative text; *(iii)* selecting, relating, organizing and interpreting information, facts, opinions and arguments in defense of a point of view; *(iv)* knowledge demonstration on the linguistic mechanisms necessary to construct the argumentation; *(v)* presentation of an intervention proposal for the problem addressed, supported by consistent arguments.

In order to avoid noise in the automatic classification process, the following processing steps were performed: *(i)* removal of special characters, numbers and dates; *(ii)* transformation of all text to lowercase; *(iii)* application of morphological markers (POS tagging) using the *nlpnet* library (Fonseca, Rosa, 2013); *(iv)* inflection of the tokens by means of stemming using the NLTK library (Bird, Klein, Loper, 2009) and the RSLP algorithm (Huyck, Orengo 2001), specific for the Portuguese language; and *(v)* segmentation (tokenization) by words, sentences and paragraphs. Beyond these steps, only the essays with more than fifty characters and whose scores available in all competencies were considered. In total, 4,317 essays were collected, dating from 2007 to 2018.

During ENEM essay evaluation, two reviewers assigned scores ranging from 0 to 200, in intervals of 50, for each of the five competencies that make up the evaluation model. Score 0 (zero) indicates that the text author does not demonstrate mastery over the competence in question. In contrast, score 200 indicates that the author demonstrates mastery over competence. If there is a difference of 100 points between the scores given by the two reviewers, the essay is analyzed by a third one. If the discrepancy persists, a group of three reviewers (Gonçalves, 2011) will evaluate the essay.

4.3 Class Balancing, Training and Testing

The corpus used for the experiments presents an imbalanced number of essays per grade that can negatively affect the classifier efficiency. Figure 4 shows the proportion of scores given for each category in Competence 5. Each bar refers to a score, with 0 referring to the lowest grade and 2 to the highest grade. This problem was approached by applying the balancing techniques described in Section 2.4. They modified the dataset used for training the models by removing samples of the majority classes (undersampling) until the distribution becomes uniform.

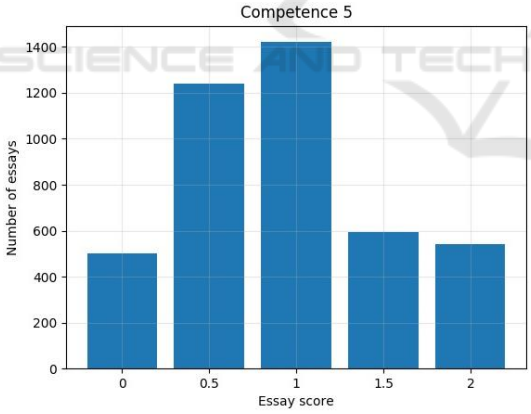

Figure 4: Class distribution of Competence 5.

For training, a matrix containing 4,317 rows and 623 columns was created, where each row represents an essay and each column represents a feature, with the last one containing the score assigned by the human evaluator for the second or fifth competence. Then, to measure the performance of each statistical model, we employed a slightly modified version of the *k*-fold cross-validation strategy (Geron, 2017).

Originally, each cycle of the *k*-fold approach splits the available data into two disjoint subsets: a test set, containing a fraction that corresponds to each possible set of the *k-1* splits, and a restricted training set containing the remaining split from the full dataset. Our modification involves the application of the balancing methods on the restricted training set while leaving the test set unchanged, so as to adequately measure their impact on the models. Moreover, we followed a stratified sampling approach, where the distribution of each class, present in the full training data, is maintained in the two subsets. Using this methodology, we guarantee that the test sets possess the same characteristics as the data that will be input into the model in a deployment environment.

Finally, each model was trained on the balanced restricted training set using the implementations provided by the scikit-learn Python library (Geron, 2017). For this work, we did not follow a systematic approach for optimizing the hyperparameters of the models due to limited resources; thus, we generally kept their default values, only changing them slightly according to emp irical observations. After training, we applied the model to the test set of the respective fold, storing their prediction and the expected score for each essay. If the model was a regressor, we rounded and scaled the predicted value to match the previously defined classes. Then, we extracted a set of metrics from the resulting data, summarizing the performance of each model and balancer; our findings are presented in the next section.

5 RESULTS

This section describes the significance of the results obtained from our experiments. It is expected that a good model-balancer pair should display a similar predictive capacity in every single class, regardless of how many examples of that class are available for training. In order to demonstrate the techniques utilized and present the results, in this article we will use Competences 5 as instance.

Considering previous results obtained with algorithms and balancing techniques in ENEM Competence 1 analyses, we chose the RUS approach to perform corpus balancing. Regarding machine learning algorithms, we chose two classification algorithms: SVC and GBT, which obtained the best results in previous work in Competence 1.

The second set of results aims at producing an overview on the performance for each pair model-balancer in Competence 5. Results are shown in Figure 5. The confusion matrices demonstrate that

argumentative features provide a marked difference in the performance of both classifiers for Competence 5. The most significant improvement was observed in the highest score (2.0) with the GBT classifier, where the true positive rate went from 0.21 to 0.48; this gain is especially remarkable if one considers the relatively small number of essays rated with this score. Additionally, in both cases, the coloring of the confusion matrices started displaying a diagonal shape when the five argumentative features were included in the model.

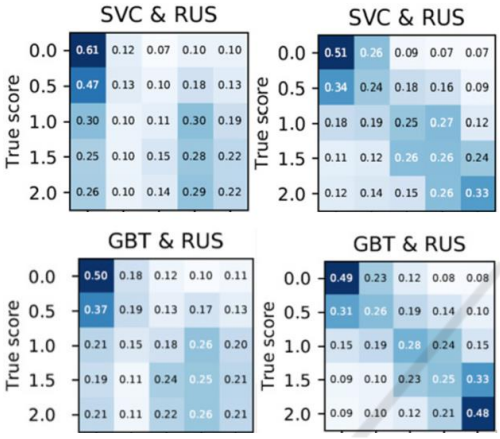

Figure 5: Normalized correlation matrices for the Competence 5 without and with argumentative features.

Brazil's ENEM regulation defines a threshold of 50 points for inter-rater disagreement for each competence, considering a maximum score of 200 points. The scale adopted for our corpus range from score 0 to 2 with the nearest adjusted threshold of 0.5. Therefore, predictions whose absolute deviation from the true score is less than 0.5 may still be considered accurate. Considering this fact, we produced a set of charts which aims at representing the current quality of each model-balancer combination. In these charts, the vertical axis represents the *relaxed* ratio of correct predictions for each class, which is equivalent to the expression $y_i = C_{i,i} + C_{i,i-1} + C_{i,i+1}$, where C indicates the cells of the normalized confusion matrix shown in Figure 6. In subtitles, lines named *With* represent accuracy including argumentative features, and lines called *Without* represent accuracy lacking argumentative features. It can be seen that the accuracy measured for Competence 5. The accuracy measured by the set of features including the arguments was much better than without them. The best accuracy becomes even more evident when we analyze the results obtained for scores 1 and 1.5. This means that a small set of only five features had

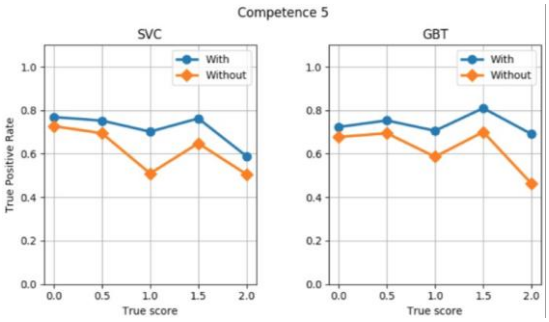

Figure 6: Relaxed ratio of correct predictions.

a good impact on a set of over 600 features.

Explanations on the causes are described in the following section.

6 DISCUSSION

Ghosh et al (2016), Nguyen and Litman (2018) have shown that the use of argumentative structures is useful for providing resources for automatic scoring systems. Our results agree with the effectivity of using argument mining incorporated into the scoring system of essays in Brazilian Portuguese.

The main criterion to evaluate Competence 5 is to check whether or not there is an intervention proposal, which comes against the new features introduced. In the set of five features there is a specific one that is to check the existence or not of an intervention proposal. In this case, the gain was very relevant, as shown by the results presented in the previous section.

We have shown that a small set of features created by following formal procedures can have very relevant results in certain domains. Only five features in a context of over 600 of them produced a highly relevant effect.

7 CONCLUSIONS AND FUTURE WORK

The main contributions of this paper are: *(i)* the identification of five new argumentative features, particularly the existence of an intervention proposal that can enrich significantly the learning process of argumentation structure, and *(ii)* the creation of an annotated corpus useful for research on argumentative structure.

We expect that as research in argumentation mining advances, prediction models will be more

accurate and argumentative features can further improve the results of automatic essay scoring. Also, it is possible for these systems to provide feedback in addition to the score. We also intend to increase the amount of texts to mark. For this experiment, we used 50 essays. We believe that with a bigger amount of tagged texts results could be improved.

REFERENCES

Almeida Junior, C. R. C. de, Spalenza, M. A., Oliveira, E., 2017. Proposal of a System for Automatic Evaluation of ENEM Essays Using Machine Learning Techniques and Natural Language Processing. *Computer on the Beach - COTB*. Florianópolis: Univali. In Portuguese.

Batista, G., Prati, R.C. and Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor*. Newslett., 6(1):20–29, 2004.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O′Reilly.

BRASIL. (2018). Redação 2018: Cartilha do Participante. http://download.inep.gov.br/educacao_basica/enem/guia_participante/2018/manual_de_redacao_do_enem_2018.pdf

Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P., 2011. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Dias da Silva, B. C., 1996. *A face tecnológica dos estudos da linguagem: o processamento automático das línguas naturais*. UNESP.

Florão, M., 2017. Como a Computação Cognitiva pode Revolucionar a Justiça Brasileira. *Computação Cognitiva e a Humanização Das Máquinas*. https://www.prodemge.gov.br/images/com_arismartbook/download/19/revista_17.pdf

Fonseca, E. R., Rosa, J. L. G., 2013. A two-step convolutional neural network approach for semantic role labeling. *IJCNN*, 1–7.

Friedman, J. H., 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5):1189–1232.

Geron, A., 2017. *Hands–On Machine Learning with Scikit–Learn and TensorFlow*. O′Reilly.

Ghosh, D., Khanam, A., Han, Y., Muresan, S., 2016. Coarse-grained Argumentation Features for Scoring Persuasive Essays. *Proceedings of the 54th Annual Meeting of the ACL*, *2*, 549–554.

Gonçalves, F. de C., 2011. *Language and genre in argumentative-dissertational essays: a systemic functional approach* (Pontifical Catholic University of São Paulo).

Haendchen Filho, A., Sartori, S., Prado, H. A., Ferneda, E., Koehntopp, P. I. (2019) An Evaluation Model for Dynamic Motivational Analysis. *ICEIS*, 1. p. 446-453.

He, H., Bai, Y., Garcia, E. A., Li, S., 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *IEEE IJCNN*, 1322–1328.

Huyck, C., Orengo, V., 2001. A Stemming Algorithm for the Portuguese Language. *8th SPIRE*, *1*, 183–193.

Johnson, R. H., Blair, J. A., 1994. *Logical Self-Defense*. International Debate Education Association.

Knerr, S., Personnaz, L., Dreyfus, G., 1990. Single-layer learning revisited: a stepwise procedure for building and training a neural network. In *Neurocomputing* Springer, p. 41–50.

Krippendorff, K., 2004. Measuring the Reliability of Qualitative Text Analysis Data. *Quality & Quantity*, *38*(6):787–800.

Lira, M. M. R. de, 2009. *Alfabetização Científica e Argumentação Escrita nas Aulas de Ciências Naturais: Pontos e Contrapontos.* UFPE.

Nguyen, H., Litman, D., 2015. Extracting Argument and Domain Words for Identifying Argument Components in Texts. *Workshop on Argumentation Mining*, *2.*, 22–28.

Nguyen, H., Litman, D., 2016. Improving Argument Mining in Student Essays by Learning and Exploiting Argument Indicators versus Essay Topics. *FLAIRS*, *29*, 485–490. AAAI Press.

Nguyen, H. V., Litman, D. J., 2018. Argument mining for improving the automated scoring of persuasive essays. *32nd AAAI*, 5892–5899.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., Napolitano, A., 2008. Building useful models from imbalanced data with sampling and boosting. *FLAIRS-21*, 306–311.

Stab, C., Gurevych, I., 2014. Identifying Argumentative Discourse Structures in Persuasive Essays. *EMNLP*, 46–56.

Stab, C., Gurevych, I., 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics*, *43*(3):619–659.

Stede, M., Schneider, J., 2018. Argumentation Mining. *Synthesis Lectures on Human Language Technologies*, 11(2):1–191.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J., 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. *13th Conference of the EACL*, 102–107.

Walton, D., 2009. Argumentation in Artificial Intelligence. In G. Simari & I. Rahwan (Eds.), *Argumentation in Artificial Intelligence*.

Yap, B. W. et al, 2014. An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. *1st Int. Conf. on Advanced Data and Information Engineering*, 285.