Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
http://br-ie.org/pub/index.php/rbie

Submission: 26/May/2018; 1st round notif.: 15/Sept/2018; New version: 10/Jan/2019; 2nd round notif.: 06/June/2019
Camera ready: 01/Aug/2019; Edition review: 01/Sept/2019; Available online: 01/Sept/2019; Published: 01/Sept/2019

Off-Topic Essay Detection:
A comparative study on the Portuguese language

Guilherme Passero
Laboratory of Applied Intelligence
University of Vale do Itajaí (UNIVALI)
guilherme.passer@gmail.com

Rafael Ferreira
Informatics Center
Federal Rural University of Pernambuco (UFRPE)
rafael.mello@ufrpe.br

Rudimar Luís Scaranto Dazzi
Laboratory of Applied Intelligence
University of Vale do Itajaí (UNIVALI)
rudimar@univali.br

Abstract

Advances in automated essay grading over the last sixty years enabled its application in real scenarios, such as classrooms and high-stakes testing. The recognition of off-topic essays is one of the tasks addressed in automated essay grading. An essay is regarded as off-topic when the student does not develop the expected prompt-related concepts, sometimes purposely. Off-topic essays may receive a zero score in high-stake tests. An off-topic essay detection mechanism may be used in parallel or embedded in an automated essay grading system to improve its performance. In this context, the main goal of this study is to evaluate the existing approaches for automated off-topic essay detection. A previous systematic review of the literature showed some deficiencies in the state of the art, including: the low accuracy of current approaches, the use of artificial validation sets, and the lack of studies focused on the Portuguese language. In this study, the approaches found in the literature, originally proposed for the English language, were adapted for the Portuguese language and compared in an experiment using a public corpus of 2164 essays related to 111 prompts. The experiment used a set of artificial off-topic examples and the best performing algorithm achieved higher accuracy than that found in the literature for the English language (96.76% vs. 94.75%). The results presented suggest the application of off-topic essay detection mechanisms in the Brazilian educational context in order to benefit the student, with computer generated feedback, and educational institutions, regarding automated essay grading. Some suggestions for future research are presented, including the need to address the task of off-topic essay detection as a multiclass problem, and to reproduce the experiment with a larger and more representative set of real off-topic essay examples.

Keywords: *Natural Language Processing, Semantic analysis, Text classification, Automated essay evaluation.*

Cite as: Passero, G., Ferreira, R. & Dazzi, R. (2019). Off-Topic Essay Detection – A comparative study on the Portuguese language. Brazilian Journal of Computers in Education (Revista Brasileira de Informática na Educação - RBIE), 27(3), 177-190. DOI: 10.5753/RBIE.2019.27.03.176

1 Introduction

Writing is widely used as an instrument for assessing student learning. In essay writing assessments, the student is asked to produce a descriptive, narrative or argumentative text, the analysis of which is not a trivial task. The process of manually grading essays presents difficulties related to time, cost, reliability and subjectivity of the evaluator. These difficulties have motivated researches aimed to automate the essay grading process (Dikli, 2006).

Advances in automated essay grading over the last sixty years enabled its application in real scenarios, such as classrooms and high-stakes testing. The pioneering research of Ellis Page, published in the 1960s, presented an approach to the problem which applied multiple linear regression with textual surface features (e.g. essay length in words, number of commas, number of prepositions etc.) (Hearst, 2000). In a comparative study, Page found a strong correlation between the computer predicted scores and the scores assigned by the teachers (Page, 1968). The progress in many fields of science, such as Artificial Intelligence and Natural Language Processing, has supported the development of many tools for automated writing analysis. Some of these tools has been successfully applied together with human graders in large-scale high-stakes testing, such as the American College Testing (ACT), the Graduate Management Admissions Test (GMAT), and the Test of English as a Foreign Language (TOEFL) (Zupanc & Bosnić, 2017).

Despite the promising results, the vulnerability to fraud of automated essay grading systems is still criticized in the scientific community (Higgins & Heilman, 2014). This problem has motivated researches to increase the reliability and performance of automated essay grading, and this includes those focused on automated off-topic essay detection, which is focus of this study.

An essay is regarded as off-topic when the student does not develop the expected prompt-related concepts, sometimes purposely. Such essay may receive a zero score in high-stakes tests. The current state of art of automated off-topic essay detection shows some deficiencies, including: the low accuracy of current tested approaches, the use of artificial validation sets, and the lack of studies focused on the Portuguese language (Passero, Ferreira, Haendchen Filho, & Dazzi, 2017). Thus, it can be seen that the automated detection of off-topic essays is an ongoing task.

Is this context, the main goal of this study is to evaluate and compare the existing approaches for automated off-topic essay detection. A comparative study was carried out on adapted versions of the existing approaches – regarding the Portuguese language and binary classification – with a Brazilian public corpus of 2,164 essays. The results obtained were discussed and compared to the literature.

This paper is structured as follows. Section 2 presents an introduction on essay assessment and off-topic essay detection. In Section 3, we describe the method and material applied in the comparative study, regarding the adaptations of the existing approaches for off-topic essay detection and the validation method used. In Section 4, we present and discuss the results found. Finally, in Section 5, we present some final considerations.

2 Essay Assessment

Writing skills are essential for school, academic and professional success (Wilson & Andrada, 2016). Essays are used in the classroom to support the development of reading and writing skills, and to assess student learning. Despite the relevance of writing skills in basic education, many Brazilian schools have not been able to stimulate this capacity adequately, as seen in the results of large-scale writing assessments performed over the last years.

In a Brazilian study, 1,500 essays submitted to a vestibular in 1978 were analyzed in detail, and only 116 (7.7%) were found to be free of coherence, cohesion, prompt-adherence and other textuality-related issues (Rocco, 2011). Also, 765 (51%) of the essays presented prompt-adherence issues. Of these, 217 (14.6%) was found to be completely off-topic and 584 (36.9%) covered the expected topic only partially.

In ENEM – *Exame Nacional do Ensino Médio* – 2016, a national exam to measure Brazilian students' performance, the submitted essays received an average grade of 543, where only 77 out of 6 million essays reached the maximum score (1000) and 55,869 (less than 1%) scored between 901 and 999 (Brasil, 2016). The rate of off-topic essays was 0.8% in the ENEM 2016 and 4.5% in the ENEM 2014.

In 2011, a national assessment of the writing of American students was carried out, where it was found that only a quarter of these students reach the level of proficiency expected for their school year (National Center for Education Statistics, 2012). Thus, it can be seen that the lack of writing proficiency is not an issue specific to Brazilian students.

An automated essay grading system may help both to reduce the human effort needed in student writing assessment and to generate feedback on essay quality, which can be used to improve the student writing skill (Wilson & Andrada, 2016). Off-topic essay detection plays a role in automated essay grading, since it helps the computer to identify essays that, on purpose or not, do not meet the presented prompt.

2.1 Off-Topic Essay Detection

An essay may be regarded as off-topic when it does not develop concepts related to the proposed theme. These essays can be classified as of two major types (Higgins, Burstein, & Attali, 2006):

- **Unexpected Topic:** possibly well-written essays that do not address the expected topic;
- **Bad-Faith:** essays that mainly consist of text copied from the prompt or with irrelevant musings, such as chunks of text unrelated to the topic.

This study addresses the detection of essays from unexpected prompt, which can be seen as a task of analyzing the textual similarity between the essay and the prompt statement (Higgins et al., 2006). In the existing literature, off-topic essay detection has been performed by applying techniques of natural language processing, semantic analysis and linguistics features, such as essay length, organization, and sentence variety (Passero et al., 2017).

A previous systematic review of the literature identified five articles, published between 2006 and 2016, which present one or more approaches for off-topic essay detection (Passero et al., 2017). In this comparative study, these approaches were considered to represent the state of the art, together with the approaches presented by Klebanov, Flor, & Gyawali (2016) and Rei & Cummins (2016), which were found to be relevant and adaptable to the task of off-topic essay detection.

Table 1, adapted from the results of Passero et al. (2017), describes the existing approaches for off-topic essay detection that were included in this comparative study. The study of Chen & Zhang (2016), although described in Passero et al. (2017), was not included in Table 1 because it did not present a novel approach.

Table 1: Approaches for off-topic essay detection found in the literature.

Source	Description
(Higgins et al., 2006)	Applies Content Vector Analysis (CVA)[1] to quantify the degree of similarity between the essay and the prompt with a variant of the *tf*idf* weighting scheme. Three models were tested for unexpected-topic detection: (A) based on the highest CVA similarity between the essay and other essays from the same prompt, and the CVA similarity between the essay and the prompt; (B) based on the frequency of words in essays from the same prompt compared to essays from other prompts; and (C) compares the similarity between the essay and its prompt to the similarity between other essays and the same prompt, and checks whether the similarity score for the essay is among the highest.
(Louis & Higgins, 2010)	Improves model C from Higgins et al. (2006), with focus on short prompts, by expanding prompt texts using: inflected forms, word association norms, WordNet synonyms, and distributionally similar words.
(Li & Yan, 2012)	Applies linear regression with an SVM model trained with two features: the proportion of prompt keywords and their similar words present in the essay; and the CVA similarity between an essay and the prompt.
(Persing & Ng, 2014)	Applies linear regression with an SVM model trained with seven groups of features: similarity between the essay and the prompt using Random Indexing (RI); presence of the most relevant 10,000 n-grams; RI similarity between the essay and each group of manually defined keywords focused on thesis clarity; RI similarity between the essay and groups of manually defined keywords focused on prompt adherence; 1000-dimensional vector representation of the essay using Latent Dirichlet Allocation (LDA); 10 features summarizing a 100-dimensional vector representation of the essay using LDA and manually defined weights; presence of thesis clarity errors.
(Klebanov et al., 2016)	Estimates the topicality of each word in the essay by comparing its occurrence in essays from the same prompt to essays from other prompts. Three techniques were applied to achieve that: significance test of Lin & Hovy (apud Klebanov et al., 2016); model B from Higgins et al. (2006); and a novel simple approach which generates a binary value – 1 if the word is more frequent in essays from the same prompt, 0 otherwise. After that, the topicality index of the essay's words is aggregated into an overall topicality score.
(Rei & Cummins, 2016)	Measures the similarity between the essay and the prompt using the cosine of the angle between their vector representation. Four techniques of vector representation were evaluated: (A) CVA; (B) Word2Vec CBOW model; (C) an improved version of B with *idf* weighting; (D) *Skip-Thoughts* neural network model; a novel approach named *Weighted-Embeddings*, an improved version of B which uses a *Skip-Thoughts*-inspired neural network for word weighting.

[1] Content Vector Analysis is a vector-based similarity measure from Information Retrieval.

3 Method and Material

In this section we describe the adaptations performed in the existing approaches for off-topic essay detection in order to enable them for the Portuguese language and the binary classification task of off-topic essays. Also, we describe the research corpus and the validation method used in the comparative study.

3.1 Adaptations of the Existing Approaches

The existing approaches for off-topic essay detection found in the literature were adapted to the Portuguese language and to binary classification. The implemented algorithms are hereafter referred to by the acronym presented in Table 2.

Table 2: Identification of the implemented algorithms.

Source	**Variant**	**Algorithm Identification (Acronym)**
(Higgins et al., 2006)	Model A	HBA-A
	Model B	HBA-B
	Model C_{UT}	HBA-C
(Louis & Higgins, 2010)	Inflected forms	LH-I
	Synonyms	LH-S
	Word association norms	LH-W
	Inflected forms + Word association norms	LH-IW
(Li & Yan, 2012)	-	LY
(Persing & Ng, 2014)	-	PN
(Chen & Zhang, 2016)	-	HBA-C
(Klebanov et al., 2016)	Significance test of topicality index	KFG-A
	Model B	HBA-B
	Proposal	KFG-B
(Rei & Cummins, 2016)	CVA	RC-A
	Word2Vec CBOW	RC-B
	Word2Vec CBOW + IDF	RC-C
	Weighted-Embeddings	RC-D

The baseline algorithms presented in Persing & Ng (2014) and Klebanov et al. (2016) were not included in this study, since the approaches proposed by these authors extend or surpass these algorithms. The technique of prompt expansion by distributionally similar words used by Louis & Higgins (2010) was not implemented because it requires a textual dependencies parser, which is still incipient for the Portuguese language. Also, the variant based on IDF-Embeddings presented by Rei & Cummins (2016) was not considered in this comparative study due to its high complexity of implementation and for performing worse than all other models in the original experiment.

All the implemented algorithms had some adaptation to enable the comparative study. These adaptations were mainly due to the difference in the natural language targeted by the studies: this is focused on the Portuguese language, while the original approaches were applied to the English language. Table 3 presents the corpora used in the implementation of the algorithms. These corpora are external resources representing the Portuguese language and, while adapting the algorithms, substituted similar sets from English language which were used in the original approaches.

Table 3: Corpora used in the adaptation of the algorithms for Portuguese language.

Identification	Description	Source
WIKIPÉDIA-PT	Collection of articles from the Wikipedia in the Portuguese language.	https://dumps.wikimedia.org/ptwiki/ (version of march/2017)
PORTAL-G1[2]	Collection of news extracted from the Brazilian Portal G1 website.	(Hartmann, 2016)

The following sections detail the implementation of the algorithms, which was performed using the Python programming language in version 3.5 and the NLTK library in version 3.2.2.

3.1.1 HBA

The HBA-A, HBA-B and HBA-C algorithms produce continuous values to represent the relevance of an essay to the topic. In previous work, threshold values were defined empirically to indicate the range of essays which were off-topic. The threshold values considered as optimal in the original papers may not present the best result in another corpus of research. Thus, the HBA-A, HBA-B and HBA-C algorithms were adapted to induce the threshold values from a training set with on- and off-topic essays. The linear SVM algorithm was chosen to treat this problem, since it allows to find the threshold value that produces the lowest overall error rate. Another advantage of the linear SVM algorithm in this context is the possibility of reducing the rate of false positives (or false negatives) by tuning the class weight parameter (Higgins et al., 2006).

3.1.2 LH

The LH-I, LH-S, LH-W and LH-IW algorithms are an extension of HBA-C focused on short prompt texts. These algorithms apply techniques of expanding the prompt using derivations, synonyms and associated words. The prompts of the research corpus, with a mean of 200 ± 38 words, are much larger than those used in the Louis & Higgins (2010) experiment (mean of 9 to 13 words). Even so, these algorithms were implemented to verify the effect on results with respect to the original HBA-C approach.

A generator of morphological variants of a word was implemented for the algorithm LH-I using a set of prefixes and suffixes of the Portuguese language, found in Marino (1980). A total of 20 prefixes of Greek origin (e.g, *ana*, *anti*, *hiper* and *meta*), 44 prefixes of Latin origin (e.g. *des*, *dis*, *inter* and *pré*), 10 augmentative nominal suffixes (e.g. *ão*, *alho* and *eirão*) 10 diminutive suffixes (e.g. *inho*, *zinho* and *ejo*) and 49 other suffixes that allow to form words in a different class (e.g. *ar*, that transforms the noun "*escola*" in the adjective "*escolar*"). Thus, for each word in a statement, many derivations (64 prefixes * 69 suffixes resulting in 4,419 variants) were generated. Most of these variants are invalid words, so a dictionary based on tokens with more than five occurrences in the WIKIPEDIA-PT corpus was queried to remove non-existent words.

The LH-S algorithm used the wordnet instance representing the Portuguese language in the Open Multilingual Wordnet project, the OpenWN-PT, while in the original work an English

[2] This corpus was used by Solo Queue (Hartmann, 2016) semantic analysis approach, winner of the workshop ASSIN-PROPOR 2016, and was granted by the author for the purpose of this research.

language instance was used. The OpenWN-PT was queried to find the synonyms of the words of the original prompt, then the synonyms found were added to the expanded version of the prompt. The words organized in this wordnet are lemmatized, while the words in the original texts of the prompt statements are in the inflected form. Therefore, the CoGrOO open source library was used to lemmatize the queried words.

The set of associated words used in the original approach of the LH-W algorithm does not exist in the Portuguese language. Thus, a similar set of associated words called PAPEL (*Palavras Associadas Porto Editora* – Linguateca) was used instead.

In tests, it could be verified that the algorithms created for expansion of prompt texts often returned a very large number of variants. This issue was addressed as in the original study of Louis & Higgins (2010), where a reduced weight was applied to the expansion words.

3.1.3 LY

The algorithm LY is a simple technique that consists of the comparison of the essay and the prompt statement using the cosine of the angle of the vectors generated by CVA. In the original approach, the computed similarity index was applied to the task of automated essay grading using linear SVR (regression). In this study, the LY algorithm uses a linear SVM classifier, a modification that aims to adapt the original approach to a binary classification task (label an essay as on- or off-topic).

In the previous work, a set of prompt keywords was used. The method of extracting or annotating keywords was not made explicit by the authors in their article[3]. Thus, in the implemented algorithm, all the content words of the prompt were used in the comparison with the essays, that is, all words except empty words (stopwords) were considered.

3.1.4 PN

In a previous study, Persing & Ng (2014) tried to estimate the prompt adherence of an essay using linear SVR (regression). In this research, the machine learning algorithm used was linear SVM for the binary classification task. Some features were excluded from this comparative study for requiring information not available in the research corpus or because they require manual annotation of the corpus, which prevents the automation of the process: thesis clarity keywords; prompt adherence keywords; manually annotated LDA topics; and predicted thesis clarity errors. Thus, the features considered in this study were n-grams and similarity between essay and prompt using the distributional semantic analysis models RI and LDA.

While in the original approach the English Gigaword corpus was used in the creation of the RI and LDA distributional models, in this adaptation we used a set of texts of the same genre (news) PORTAL-G1. The pre-processing techniques adopted in the original experiment were reproduced in the research corpus essays and in the PORTAL-G1 corpus. These techniques included the conversion of texts to lowercase and lemmatization using the CoGrOO library.

[3] We attempted to contact Li & Yan (2012) to obtain more details on their keyword extraction or annotation method, but we have not received a response.

3.1.5 KFG

The approaches presented by Klebanov et al. (2016) are based on building a dictionary with the words present in the essays of a certain prompt and an index of topicality for each word. With this dictionary, the topicality of an essay can be summarized from the indices referring to its words. In the original study, the dictionaries were generated with 500 essays per prompt. In the context of this research, there is a limited set of 20 essays per prompt, thus all of them were used in the construction of the dictionaries.

The authors evaluated four techniques for summarizing words' topicality indices in a single index referring to an essay, considering: all words (tokens) or unique words (types); and length-scaled or not length-scaled mean of words' indices. The KFG-A and KFG-B algorithms were implemented using the length-scale summarization of single words because this setting optimized the results in the original experiment.

In the original study, the generated topicality indices were used to predict a holistic score for essays using linear regression. Since this research is aimed to classify an essay as on- or off-topic, a linear SVM classifier was adopted in the place of regression.

For the implementation of the KFG-A algorithm, which uses a test of significance, the binary model proposed by Klebanov et al. (2016) was adopted, since it was observed as the optimal model in the original experiment, where the words that appear most frequently in essays with a significance limit of 0.05.

3.1.6 RC

The RC- * algorithms consist of converting the text of the prompt and the essay into a vector representation and then obtain an index of similarity between these texts using the cosine of the angle. In the previous work, the sentences of an essay were checked individually in order to highlight specific parts of the essay that were off-topic. In this research, the method of extracting and comparing the vectors was not modified, but the scope of the comparison was changed, from the sentence to the full text of the essay.

The RC-B, RC-C and RC-E algorithms depend on a Word2Vec model for vectoring texts. In this study, a similarity analysis model with Word2Vec was created using the corpora WIKIPÉDIA-PT and PORTAL-G1, whereas in the original work the British National Corpus text collection was used. The authors did not specify the preprocessing techniques used in their experiment, so we chose the procedure adopted in the Solo Queue approach (Hartmann, 2016), which also includes a Word2Vec model for semantic analysis.

The RC-* algorithms produce a ranking of the prompts most likely related to the essay, as in the HBA-C algorithm. Thus, a linear SVM classifier was adopted to identify the best ranking position to classify an essay as on- or off-topic.

3.1.7 Content Vector Analysis

In this study, the implementation of the CVA – Content Vector Analysis – algorithm was based on the corpora WIKIPÉDIA-PT and PORTAL-G1 to determine the frequency of terms in documents, necessary for the calculation of the *tf*idf* weight, while the HBA- * and LH- * algorithms used the collection of newspaper texts TIPSTER and the RC-A algorithm used the British National Corpus. This adaptation was necessary because the collections of original texts were in the English language. The two sets - WIKIPÉDIA-PT and PORTAL-G1 - were selected to replace the original collections for also having a varied and large amount of texts, which helps ensure the stability of the frequencies extracted from terms in documents (Higgins et al., 2006).

3.2 Validation

The research corpus contains 2,164 essays related to 111 prompts (~20 essays per prompt) and was obtained from a Brazilian website[4] focused on helping students with improving their writing skills. The pages of the website were parsed using an own web crawler and the essays were converted to a structured file in the XML format with each essay's URL, original text, corrected text, final score, criteria's score, the prompt and comments of the rater. The web crawler and the structured file with the research corpus were made publicly available[5].

We asked two reviewers to classify the essays of the corpus as on-topic or off-topic aiming to build a corpus with real examples, however the output set of off-topic essays was too small (only 12 out of the 2,164 essays). Since we could not build a representative set with real examples of off-topic essays from the research corpus, the performance of the algorithms was evaluated in the research corpus using sets of artificial examples created using the strategy observed in the literature for artificial positive (off-topic) examples generation. The strategy consists in using random essays as off-topic essay: for each set of *N* essays of a prompt (negative or on-topic examples), *N* essays are randomly selected from other prompts (positive or off-topic examples).

The experiment was performed using a cross-validation strategy by prompt. For each prompt with *N* essays, on-topic and off-topic examples included, *N* validation cycles were executed, where for each cycle an essay of the prompt composed the test set and the other essays composed the training set (*leave-one-out* method). The statements from the other prompts were used as a development set to make the experiments compatible with all the evaluated algorithms. Thus, in each validation cycle, the ability of an algorithm to correctly classify an essay as a positive or negative example was tested based on the prompts and other essays of the research corpus.

The performance of the algorithms was measured by the overall rate of correct classifications (accuracy), precision, recall, false positive rate and false negative rate. These measures were used because they are present in the literature related to off-topic essay detection, as presented by Passero et al. (2017), and, therefore, aim to facilitate the comparison of the results achieved with the one found in other studies.

4 Results and Discussion

The performance results obtained for the adapted off-topic essay detection approaches in the research corpus are presented in Table 4 regarding accuracy, precision, recall, false positive rate (% FP) and false negative rate (% FN). A comparison of the results found with previous studies is presented in Appendix A.

[4] *UOL Educação – Banco de Redações* <https://educacao.uol.com.br/bancoderedacoes/>.

[5] <https://github.com/gpassero/uol-redacoes-xml>.

Table 4: Performance of the algorithms validated in the research corpus.

Algorithm	% Accuracy	% Precision	% Recall	% FP	% FN
HBA-A	90.31	88.79	92.26	11.65	7.74
HBA-B	89.86	88.52	91.61	11.88	8.39
HBA-C	83.06	89.66	74.74	8.62	25.26
LH-I	82.62	89.01	74.45	9.20	25.55
LH-S	82.60	89.41	73.95	8.76	26.05
LH-W	81.13	87.78	73.32	10.07	27.68
LH-IW	81.62	87.93	73.30	10.07	27.70
LY	86.84	83.85	91.24	17.57	8.76
PN	85.76	88.63	82.06	10.53	17.94
KFG-A	**96.76**	**95.84**	**97.76**	**4.24**	**2.24**
KFG-B	94.62	94.68	94.55	5.31	5.45
RC-A	81.36	77.93	87.51	24.79	12.49
RC-B	83.13	88.39	76.28	10.02	23.72
RC-C	84.30	89.61	77.59	8.99	22.41
RC-D	88.37	92.69	83.32	6.57	16.68

In the previous experiment carried out by Higgins, Burstein & Attali (2006), the original algorithms HBA-A and HBA-B and HBA-C achieved false positive rates of 4.7% to 6.9% and false negative rates of 16.8 % to 38%. The results obtained in this experiment, regarding the FP and FN measures, differ from those found in the previous study designed for the English language. This occurred because the original study optimized the needed threshold for the specific corpus of interest to achieve a low rate of false positives. By using the same approach of optimizing the decision boundary threshold[6] instead of using the SVM learning algorithm, the results obtained were close to the ones observed in the previous study: FP 6.33% and FN 27.21%.

Since the data set of the study by Higgins, Burstein & Attali (2006) was balanced, it can be estimated that the general accuracy obtained for the HBA-A, HBA-B and HBA-C algorithms was, respectively, 78.5%, 83.55% and 85.15%. The difference between the accuracies obtained in the previous study and the ones found in this research may be due to differences in the corpus of essays and the learning strategy of the decision boundaries, which in this experiment occurred with SVM and the original study optimized the threshold values of the indexes of similarity analyzing the trade-off curve between FP and FN. Still, in this research the results obtained for these algorithms present a different trend from that found in the literature: the HBA-A and HBA-B algorithms reached an accuracy higher than that found in the original study; while the HBA-C algorithm achieved lower accuracy. This difference may have occurred because the HBA-A and HBA-B algorithms use a training set with essays of the same subject, whereas the HBA-C uses as a set of varied prompt texts for training. Thus, the particularities of the corpus of this research may have allowed a better use of essays from the same prompt in a training set compared to the previous study.

[6] In the previous experiment the decision boundary was produced by the threshold value of 10, or 29% of the total number of reference prompt (34 + 1). Thus, an essay was classified as on-topic if the similarity between the essay text and the prompt was among the 10 prompts with the highest index of similarity (considering the reference prompts). The FP and FN presented measures was produced using the threshold value of 32 (29% of 111 prompts).

The algorithms LH-I, LH-S, LH-W and LH-IW are an extension of the HBA-C algorithm that uses prompt expansion techniques (e.g. addition of synonyms and associated words), so the prompt may more broadly cover the expected topic. These algorithms were developed by Louis & Higgins (2010) to improve the performance of the original HBA-C algorithm in cases where the text of the prompt is very short. The results presented in Table 4 indicate that the expansion techniques of the LH-I, LH-S, LH-W and LH-IW algorithms result in a worse performance than the original HBA-C algorithm, considering the corpus of this research and the measures used. On the other hand, in the experiment of Louis & Higgins (2010) these algorithms reached a result better than that found by Higgins, Burstein & Attali (2006) in the classification of off-topic essays, considering prompt statements with the average length of 9 to 13 words. In the corpus of this research, the prompt statements have 200 words on average. Thus, these results suggest that the LH-I, LH-S, LH-W and LH-IW algorithms are more adequate for cases where the text of the prompt has less than 200 words.

The algorithm LY can be seen as an extension of the HBA-C algorithm, since it also compares the text of the prompt and the essay using CVA. Nonetheless, this algorithm introduces a second feature: the proportion of prompt keywords present in the essay, which in this research was regarded as all the prompt words, except empty words (stopwords). In this experiment, the LY algorithm presented an improvement over the previous one – HBA-C – considering the accuracy measure, but a lower precision.

In the review of the state of the art, it has been found that only in the experiments of Higgins, Burstein & Attali (2006) (a), Louis & Higgins (2010) (b) and Chen & Zhang (2016) (c) the presented approaches were validated in a binary-class corpus, where false positive and false negative rates (a and b) or precision, recall and F-value (c) were measured. The best performance found in the literature for the task of binary classification of off-topic essays, in relation to accuracy, was obtained by Louis & Higgins (2010) in a set of essays written by advanced English-speaking students, where the original LH-W algorithm achieved an accuracy of 94.75%.

In this experiment, the KFG-A algorithm achieved the higher accuracy in performance (96.75%). Regarding previous studies applied to binary classification task of off-topic essays, the KFG-A algorithm had the highest accuracy. It is worth mentioning the original experiment for KFG-A, which was performed by Klebanov et al. (2016), has not measured the performance of this algorithm in the task of off-topic essay detection – but in the automated essay grading task.

The compared algorithms that do not require a corpus of on-topic essay are: HBA-C, LH-* and RC-*. Among these algorithms, the RC-D performed better regarding all the measures used, with an accuracy loss of 8.49% in relation to the best performing approach – KFG-A –, which depends on essays from the same prompt for training. The RC-D algorithm uses only the text of the essay and the prompt, differently from every other algorithm compared in this study, and it may be considered in scenarios that lack the additional resources required by the other approaches (set of off-topic essays and set of prompt texts).

This experimented used a corpus with at most 20 essays from the same topic for training, which is the smallest training set compared to the previous studies (~63 to 48.488). The results obtained for the KFG-A algorithm, which has used a training set of 500 essays in the previous study by Klebanov et al. (2016), when compared to the results found for the algorithms that do not require this type of training set – HBA-C, LH-* and RC-*, shows that even a small training set of on-topic essays may improve the performance of an off-topic essay classifier.

In the experiment of Chen & Zhang (2016), the original version of the HBA-C algorithm was applied in sets of real examples of off-topic essays, and the results obtained in this experiment differ in the precision, recall and F-value measures. In the previous study, the four groups analyzed were 100% accuracy, recall between 2.2% and 18.1%, and F-value between 4.4% and 30.7%. Since the validation set of the present experiment is balanced, it can be inferred an F-value equal to the accuracy. Thus, in relation to the F-value measure, this experiment presented results superior to that of Chen & Zhang (2016) for the HBA-C algorithm, as was observed in the Higgins, Burstein & Attali (2006) experiment with artificial examples.

5 Conclusions

In this paper we presented a comparative study of the existing approaches for automated off-topic essay detection. The approaches were adapted to the Portuguese language and the binary classification task, and validated in a Brazilian corpus of 2,164 essays. The presented results show that the performance of the evaluated algorithms, as measured by accuracy, ranges from 81.13% to 96.76%. The algorithm with the best accuracy in this experiment (96.76%) was KFG-A, an adapted version of the approach proposed by Klebanov et al. (2016). The results were compared to those found in the literature, where some differences were observed and discussed.

The results found suggest the development and application of automated off-topic essay detection systems in the Brazilian educational context in order to benefit the student, by generating feedback on essays, and educational institutions, by supporting the automation of the essay grading process.

We suggest for future research to validate the existing approaches for off-topic essay detection with a representative set of real examples. Also, it may be relevant to treat off-topic essay detection as a multiclass problem, thereby the performance of the proposed approaches might be evaluated considering the many types of off-topic essays, such as: well written essays that do not address the expected topic (addressed in this research); essays mainly composed of copies of the prompt text; essays that do not answer the main prompt question; and essays with purposely disconnected parts. We also suggest for future research to optimize the classifiers applied to off-topic essay detection in order to reach an ideal false positive rate (e.g. the one found in the human evaluation).

References

Brasil. (2016). ENEM 2016: Resultado Individual [ENEM 2016: Individual Result]. http://download.inep.gov.br/educacao_basica/enem/downloads/2016/apresentacao_final_resultados_2016.pdf

Chen, J., & Zhang, M. (2016). Identifying Useful Features to Detect Off-Topic Essays in Automated Scoring Without Using Topic-Specific Training Essays. *Springer Proceedings in Mathematics and Statistics*, *140*(August), 315–326. DOI:10.1007/978-3-319-19977-1 [GS Search]

Dikli, S. (2006). An Overview of Automated Scoring of Essays. *Journal Of Technology Learning And Assessment*, *5*(1). https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1640/1489 [GS Search]

Hartmann, N. S. (2016). Solo Queue at ASSIN : Combinando Abordagens Tradicionais e Emergentes [Solo Queue at ASSIN: Combining Traditional and Emerging Approaches]. In *PROPOR – International Conference on the Computational Processing of Portuguese* (p. 6). http://propor2016.di.fc.ul.pt/wp-content/uploads/2015/10/ASSIN-2016-solo-queue.pdf [GS Search]

Hearst, M. (2000). The debate on automated essay grading. *Intelligent Systems and Their Applications, IEEE*, *15*(5), 22–37. DOI:10.1109/5254.889104 [GS Search]

Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, *12*(2), 145–159. DOI:10.1017/S1351324906004189 [GS Search]

Higgins, D., & Heilman, M. (2014). Managing what we can measure: Quantifying the susceptibility of automated scoring systems to gaming behavior. *Educational Measurement: Issues and Practice*, *33*(3), 36–46. DOI:10.1111/emip.12036 [GS Search]

Klebanov, B. B., Flor, M., & Gyawali, B. (2016). Topicality-Based Indices for Essay Scoring. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 63–72. DOI:10.18653/v1/W16-0507 [GS Search]

Li, Y., & Yan, Y. (2012). An effective automated essay scoring system using support vector regression. *Proceedings - 2012 5th International Conference on Intelligent Computation Technology and Automation, ICICTA 2012*, 65–68. DOI:10.1109/ICICTA.2012.23 [GS Search]

Louis, A., & Higgins, D. (2010). Off-topic essay detection using short prompt texts. *NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, (June), 92–95. https://www.aclweb.org/anthology/W10-1013.pdf [GS Search]

Marino, E. R. (1980). *Estudos de Português para o 2º Grau [Portuguese Studies for Highschool]*. Editora do Brasil, 1st ed. São Paulo. [GS Search]

National Center for Education Statistics. (2012). The Nation’s Report Card: Writing 2011. *The Nation’s Report Card: Writing 2011*. https://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf [GS Search]

Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education*, *14*(2), 210–225. DOI:10.1007/BF01419938 [GS Search]

Passero, G., Ferreira, R., Haendchen Filho, A., & Dazzi, R. (2017). Off-Topic Essay Detection: A Systematic Review. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)* (Vol. 28, p. 51). DOI:10.5753/cbie.sbie.2017.51 [GS Search]

Persing, I., & Ng, V. (2014). Modeling Prompt Adherence in Student Essays. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, (June), 1534–1543. DOI:10.3115/v1/P15-1053 [GS Search]

Rei, M., & Cummins, R. (2016). Sentence Similarity Measures for Fine-Grained Estimation of Topical Relevance in Learner Essays. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 283–288. DOI:10.18653/v1/W16-0533 [GS Search]

Rocco, M. T. F. (2011). Crise na linguagem: a redação no vestibular [Crisis in language: the essay in the vestibular exam]. *Em Aberto*, 2(12). [GS Search]

Wilson, J., & Andrada, G. N. (2016). Using Automated Feedback to Improve Writing Quality: Opportunities and Challenges. In *Handbook of Research on Technology Tools for Real-World Skill Development* (pp. 678–703). Hershey: Information Science Reference. DOI:10.4018/978-1-4666-9441-5.ch026 [GS Search]

Zupanc, K., & Bosnić, Z. (2017). Automated essay evaluation with semantic analysis. *Knowledge-Based Systems, 120*, 118–132. DOI:10.1016/j.knosys.2017.01.006 [GS Search]

Appendix

A. Comparison of Our Results with Previous Studies

	Our Result (for Portuguese)					Result of Previous Study (for English)					
Algorithm	**% Acc.**	**% Prec.**	**% Recall**	**% FP**	**% FN**	**Ref.**	**% Acc.**	**% Prec.**	**% Recall**	**% FP**	**% FN**
HBA-A	90.31	88.79	92.26	11.65	7.74	Higgin et al. (2006)	78.50	-	-	5.0	38.0
HBA-B	89.86	88.52	91.61	11.88	8.39		83.55	-	-	4.7	28.2
HBA-C	83.06	89.66	74.74	8.62	25.26		85.15			6.8	22.9
						Chen & Zhang (2016)	4.4-28.6	100.0	2.2-18.1	-	-
LH-I	82.62	89.01	74.45	9.20	25.55	Louis & Higgins (2010)	91.05-94.20	-	-	2.53-6.25	9.06-11.65
LH-S	82.60	89.41	73.95	8.76	26.05		90.48-94.45	-	-	1.39-7.03	9.76-12.01
LH-W	81.13	87.78	73.32	10.07	27.68		90.85-94.75	-	-	1.47-6.33	9.02-11.97
LH-IW	81.62	87.93	73.30	10.07	27.70		91.24	-	-	6.04	11.48
Best – KFG-A	96.76	95.84	97.76	4.24	2.24	Best – Louis & Higgins (2010)	94.75	-	-	1.47	9.02