Off-Topic Essay Detection: A Systematic Review

Guilherme Passero[1,2], Rafael Ferreira[3], Aluizio Haendchen Filho[2], Rudimar Luís Scaranto Dazzi[1]

[1]Laboratory of Applied Intelligence,
University of Vale do Itajaí (UNIVALI) – Itajaí, SC – Brazil

[2]Department of Artificial Intelligence and Smart Systems,
University Center of Brusque (UNIFEBE) – Brusque, SC – Brazil

[3]Department of Statistics and Information,
Federal Rural University of Pernambuco (UFRPE) – Recife, PE – Brazil

{guilherme.passer, rafaelflmello, aluizio.h.filho}@gmail.com,
rudimar@univali.br

Abstract. *Essays are widely used for learning assessment in the educational context. Commercial solutions for automated essay scoring have shown promising results, but vulnerability to fraud is still criticized in the scientific community. An off-topic essay detection tool can be used to increase the reliability of automated essay scoring systems and to generate feedback to students. In this context, this paper presents a systematic review of the literature on automatic detection of off-topic essays. We describe the techniques and resources, the corpora and the performance of existing approaches. The results found indicate some gaps and deficiencies in the existing literature, including the need to reduce error rates and to use validation sets based on real examples of off-topic essays.*

1. Introduction

Essays are widely used for learning assessment in the educational context. An essay test evaluates the competencies developed by a student and promotes improvements in communication and expression skills. In an essay, a statement with a thematic proposal (prompt) is presented to the student, demanding the elaboration of a descriptive, narrative or argumentative textual response. The analysis of such texts is not a trivial task [Santos, Paiva and Bittencourt 2016]. The teacher may spend considerable time in evaluating an essay content, since there are many textual features to be analyzed.

Some of the commercial solutions for automated grading of student-written essays have shown feasibility for high-stakes testing. Despite the promising results reported, the vulnerability to fraud of these systems have been criticized in the scientific community [Higgins and Heilman 2014]. For instance, a well-written essay that does not address the proposed topic may receive an overestimated score from an automated grader because of linguistic features, such as text structure and surface. One solution to mitigate this problem is to use an off-topic essay detection tool alongside or embedded into the existing essay scoring systems [Higgins and Heilman 2014, Chen and Zhang

DOI: 10.5753/cbie.sbie.2017.51

2016]. A tool for automatic off-topic essay detection might also be used to generate relevant feedback to students [Higgins, Burstein and Attali 2006].

In this context, this paper presents a systematic review of the literature on automatic detection of off-topic essays. In Section 2, we present an introduction on off-topic essay detection. Section 3 deals with the search protocol applied in this study, including the research questions, data sources, search strategy and study selection. In Section 4, we present the results found and an attempt to answer the research questions. At last, in Section 5, we present some final considerations.

2. Off-topic Essay Detection

Prompt adherence is an often-used criterion in essay evaluation. An essay must develop concepts in various areas of knowledge to meet this criterion, and such concepts must be related to the proposed statement. When the prompt-adherence criterion is not met, an essay may be rated as off-topic. Off-topic essays can be regarded as of two major types [Higgins, Burstein and 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 purposely inserted chunks of text unrelated to the topic and the essay itself.

The detection of off-topic essays can be seen as a task of analyzing the closeness between the content of an essay and the prompt statement [Higgins, Burstein and Attali 2006]. Linguistic features such as essay length, organization, and sentence variety are also relevant for off-topic essay detection [Chen and Zhang 2016]. In the existing literature, off-topic essay detection has been performed by applying techniques of natural language processing which may or may not regard semantic aspects of the text.

A training set of essays written to the same prompt (most likely on-topic) might be used to improve the results of an off-topic essay detector, as in the work of Li and Yan (2012), Persing and Ng (2014) and Chen and Zhang (2016), but a topic-specific corpus is not always available in real scenarios. In some studies, a reference set of essays or prompt descriptions from previous applications is used alone to help the definition of threshold values and weights for features, which can be made empirically or by machine learning [Higgins, Burstein and Attali 2006, Louis and Higgins 2010].

3. Method

According to Kitchenham and Charters (2007), a systematic literature review (SR) is “a means of identifying, evaluating and interpreting all available research relevant to a particular research question, or topic area, or phenomenon of interest”. The most common reasons to undertake a systematic review are: summarize existing evidence on a treatment of technology; identify gaps in current research; and provide a background in order to appropriately position new research activities [Kitchenham and Charters 2007].

The aim of this SR is to provide an overview of the current research on automatic detection of off-topic essays. The SR was carried out following the guidelines

defined by Kitchenham and Charters (2007). The method applied in the SR is presented in the next sections.

3.1. Research questions

The overview of the current research on off-topic essay detection was addressed as describing the techniques and resources, the corpora and the performance measured in the existing literature. The research questions of this SR are presented in Table 1.

Table 1. Research questions

ID	Research question
Q1	What techniques and resources have been used in the existing approaches?
Q2	Which corpora have the existing approaches been tested on?
Q3	How accurate are the existing approaches?

3.2. Data sources and search strategy

The electronic bibliographic databases presented in Table 2 were used as primary sources for research papers in this SR. The data sources were selected by meeting the criteria suggested in [Silva et al. 2016]: digital libraries consolidated in the Computer Science; digital libraries that enable search by string with keywords; and digital libraries that enable online access.

Table 2. Data sources

Source	URL
ACM Digital Library	http://dl.acm.org/
IEEE Explore Digital Library	http://ieeexplore.ieee.org/
Science Direct	http://www.sciencedirect.com/
Scopus	http://www.scopus.com/
Proceedings of the Brazilian Symposium on Computers in Education (SBIE)	http://www.br-ie.org/pub/index.php/sbie
Brazilian Journal of Computers in Education (RBIE)	http://www.br-ie.org/pub/index.php/rbie

In order to define the search string, we used the terms “off topic” and “essay” to find papers on off-topic essay detection. The papers found, including its cited papers, have been analyzed to find potential terms to include in the search string. After several trials, we defined the two search strings presented in Table 3 in order to return the most relevant papers on the research topic for English and Portuguese languages.

Table 3. Search strings

English	("off-topic" OR "off topic" OR “prompt adherence”) AND "essay"
Portuguese	(“fuga ao tema” OR “adequação ao tema”) AND (“redação” OR “redações”)

The search strings were adapted for each data source, preserving its logical connectives to return only papers containing the desired terms in its title, abstract or keywords.

3.3. Paper selection

The inclusion and exclusion criteria defined for papers selections are presented in Table 4. The exclusion criterion CE3 was added because of the presence of a retracted paper on the search results, which the publisher recommended not to be considered in future studies. The approach described in the excluded paper is however extended in one of the selected papers: [Li and Yan 2012].

Table 4. Inclusion and exclusion criteria

Inclusion	Exclusion
CI1. Papers published until 18/06/2017. CI2. Papers in English and Portuguese languages. CI3. The title, abstract or keywords meet one of the search strings.	CE1. Short papers and extended abstracts. CE2. Papers unrelated to search scope. CE3. Papers retracted by the publisher.

After the search in the data sources, the parts of the retrieved papers were analyzed in this sequence: title, keywords, abstract, introduction, and conclusion. The inclusion and exclusion criteria were applied to check whether the study is directly related to the topic of this review and has the potential to answer the research questions.

4. Results

In the initial search using inclusion criteria, 14 papers were returned from the selected data sources. Nine papers were excluded from the review, five of them because of non-adherence to the scope of the research, three for duplicity and one for having been retracted by the publisher. Finally, five papers were selected for the review. The number of papers found and selected by executing the search protocol (Section 3) is presented in Table 5 broken down by data source.

The next section presents an overview on the reviewed studies. Thereon we present a comparative analysis and discuss the results in regard of the research questions.

Table 5. Number of papers retrieved and selected

Data source	Retrieved	Selected	Excluded
ACM Digital Library	1	1	0
IEEE Explore Digital Library	4	1	3
Science Direct	1	0	1
Scopus	8	4[1]	4
SBIE	0	0	0
RBIE	0	0	0
Total	**14**	**5**	**8**

4.1. Studies overview

In this section, we present an overview on the five selected papers listed in Table 6.

[1] One of the selected papers was duplicate, since it was already found IEEE's data source. So, the total number of selected papers is the sum of this column minus one, corresponding to the duplicate paper.

Table 6. List of selected papers

Reference	Title
[Higgins, Burstein and Attali 2006]	Identifying off-topic student essays without topic-specific training data
[Louis and Higgins 2010]	Off-topic essay detection using short prompt texts
[Li and Yan 2012]	An effective automated essay scoring system using support vector regression
[Persing and Ng 2014]	Modeling prompt adherence in student essays
[Chen and Zhang 2016]	Identifying useful features to detect off-topic essays in automated scoring without using topic-specific training essays

Higgins, Burstein and Attali (2006) present an approach to detect off-topic essays without the need of topic-specific training data to improve the software Criterion^{SM}, which required a training corpus of 200 to 300 labeled essays. Content Vector Analysis (CVA), a vector-based similarity measure from Information Retrieval, was applied in this study to quantify the degree of similarity between the essay text and the prompt description with a variant of the *tf***idf* weighting scheme. The following four models were tested in the study, where the first two are previous models which depend on topic-specific training data, and the latter two are the novel approaches which are based on non-topic-specific training data:

- **Model A:** based on the highest CVA similarity between an essay and other essays from the same prompt and between the essay text and the prompt description;
- **Model B:** based on the occurrence of the words of an essay in essays from the same prompt (specific rate) and across different prompts (global rate);
- **Model C_{UT} (Unexpected Topic):** compares the essay text to its prompt and a set of reference prompts using CVA and checks whether the measured similarity score for the target essay prompt is among the highest scores;
- **Model C_{BF} (Bad-faith):** predicts bad-faith essays using support vector machine (SVM) with five features: CVA similarity to the prompt; number of content words; proportion of content words not found in the prompt; ratio of word types to word tokens; and the presence of specific "direct address" markers (e.g. "hello", "thanks").

Louis and Higgins (2010) present methods to detect off-topic essays with short prompts by extending Model C_{UT} from Higgins, Burstein and Attali (2006). The authors noticed a very high rate of false-positives when applying the previous model to a set of prompts with the average word size between 9 and 13 (2,94%-9,73%), whereas prompts with 60 to 276 words in average have shown a lower rate (0,20%-0,73%). Among the evaluated methods to expand short prompts and decrease the false-positives rate, the best results were obtained by using spelling correction, inflected forms (i.e. "*friendly*" expands to "*friend*", "*friendlier*" and "*friendliness*") and word association norms (i.e. "*friendly*" expands to "*smile*", "*amiable*", "*greet*" and others). Word association norms (WAN) is a collection of 5.000 words and their associations produced by about 6.000 people in a previous study [Nelson et al. 1998 apud Louis and Higgins 2010]. The expansion of prompts using WordNet synonyms and an approach for deriving distributionally similar words (henceforth DSW) presented in [Lin 1998 apud Louis and Higgins 2010] were also tested, but a slightly lower accuracy was achieved.

Li and Yang (2012) present an approach to estimate the prompt adherence of essays as part of a system for automated essay grading. A linear regression model using SVM was trained with two features extracted from the essays: the proportion of the prompt keywords and their similar words present in the essay, and the CVA similarity between an essay and the prompt. The authors found that the predicted topic scores compared well to the essays' overall score by analyzing a chart with the relation of the predicted values and the human score.

Persing and Ng (2014) address the task of estimating the relatedness between an essay and its prompt. The authors applied SVM linear regression, creating a specific prediction model for each prompt from the research corpus. The predicted values ranged from one (completely off-topic) to four (completely on-topic). The features extracted from the datasets and used in creating machine learning models were:

- **Random Indexing:** semantic similarity between essays and the prompt, whole text and by sentence, using a distributional model (Random Indexing or RI);
- **N-grams:** presence of most relevant 10.000 uni, bi and tri-grams;
- **Thesis Clarity Keywords:** RI similarity between the essay and each group of the manually defined prompt primary and secondary keywords introduced in [Persing and Ng 2013 apud Persing and Ng 2014];
- **Prompt Adherence Keywords:** RI similarity between the essay and groups of manually defined keywords;
- **LDA Topics:** 1000-dimensional vector representation of the essay using another semantic distributional model, the Latent Dirichlet Allocation (LDA);
- **Manually Annotated LDA Topics:** 100-dimensional vector representation of the essay by applying LDA, where each dimension (or topic) had a weight set in accordance with its prompt adherence and 10 features were generated by summarizing the vectors values (estimated topic adherence) for each weight;
- **Predicted Thesis Clarity Errors:** binary features indicating the presence of predicted clarity errors, including problems related to confusing phrasing, incomplete prompt response, relevance to prompt, missing details and writer position.

Through recursive feature elimination, the authors found that the most relevant extracted features were n-grams, clarity keywords and manual LDA topics, whereas RI and automatically extracted LDA topics were regarded as of middle importance. It is worth noting that despite of the potential of classifying off-topic essays, the tested corpus had no essay classified as completely off-topic.

Chen and Zhang (2016) deals with the problem of off-topic essay detection aiming to improve E-rater® commercial solution – an automated essay scoring system. The previously described Model C_{UT} from [Higgins, Burstein and Attali 2006] was being applied by E-rater® to detect off-topic essays, and this study tried to find relevant features which could be used to improve the results. Although the paper's main focus was on the detection of off-topic essays without a topic-specific corpus, the research used an external set of essays from the same prompt to compute similarity features.

First, the authors identified features relevant to the problem. Then, these features were applied in a case study to check its reliability in the task of classifying off-topic essays. The most distinctive features according to the study analysis were: (i) essay

length in number of characters, words and sentences; (ii) ratio of the number of word types to the number of word tokens; (iii) CVA similarity between the essay and a set of training essays; (iv) organization; and (v) sentence variety. The organization feature was extracted from the essay by detecting the presence of a particular discourse element, such as introductory material, a thesis statement, supporting ideas and a conclusion.

In the reviewed literature, Chen and Zhang (2016) presented the only study to use a real set of off-topic essays of both "unexpected topic" and "bad-faith" types. Despite the perfect rate of precision (100%), a low recall rate was found (2.2-18.1%). The recall rate was not measured regarding the types of off-topic essays, what the authors considered as one of the limitations of the study.

4.2. Techniques and resources (Q1)

The first research question is about the techniques and resources used in existing literature to address the task of detecting off-topic essays. The existing approaches use many different techniques of natural language processing, semantic analysis, probabilistic estimation and machine learning. We also found that some of the reviewed papers have evaluated several approaches to achieve the best results for each of the research datasets.

The techniques used in the reviewed literature are presented in Table 7 with respect to three major types: (i) probabilistic surface analysis, referring to techniques which use text surface features and probabilistic estimations (e.g. CVA and dictionary of word occurrence on and off-topic); (ii) semantic analysis, a class for techniques that measure text similarity which are corpus-based (LDA, RI and DSW) or thesauri-based (WordNet and word association norms); and (iii) machine learning, which includes the use of linear regression with SVM to estimate the prompt adherence of an essay based on a training set of reference essays.

Table 7. Techniques used in existing literature

Reference	Probabilistic surface analysis	Semantic analysis	Machine learning
[Higgins, Burstein and Attali 2006] Models A, B and C_{UT}	Yes	No	No
[Higgins, Burstein and Attali 2006] Model C_{BF}	Yes	No	Yes
[Louis and Higgins 2010]	Yes	Yes	No
[Li and Yan 2012]	Yes	No	Yes
[Persing and Ng 2014]	Yes	Yes	Yes
[Chen and Zhang 2016]	Yes	No	No

The approaches found in the literature are presented in Table 8 regarding the type of training data used to evaluate unseen essays. The composition of training data is relevant to the deployment of an off-topic essay detection system in a real scenario, since it specifies which type of data must be available for the system to function.

Table 8. Composition of the training datasets used in existing literature

Reference	Approach	Same prompt description	Other prompt descriptions	Same prompt essays	Other prompt essays
[Higgins, Burstein and Attali 2006]	Model A	Yes	No	Yes	No
	Model B	No	No	Yes	Yes
	Model C_{UT}	Yes	Yes	No	No
	Model C_{BF}	Yes	Yes	No	Yes
[Louis and Higgins 2010]	Model C_{UT1}	Yes	Yes	No	No
[Li and Yan 2012]	Proposal	Yes	No	Yes	No
[Persing and Ng 2014]	Baseline	No	No	Yes	No
	Proposal	Yes	Yes	Yes	No
[Chen and Zhang 2016]	Model C_{UT2}	Yes	Yes	Yes	No

The approaches presented in [Louis and Higgins 2010] and [Li and Yan 2012] were respectively named as "Model C_{UT1}" and "Model C_{UT2}", once they extend Model C_{UT} from [Higgins, Burstein and Attali 2006], the first by applying techniques for short prompt expansion and the second by adding linguistic and similarity features.

4.3. Corpora (Q2)

The second research question is about the corpora which have been evaluated in existing researches for automatic off-topic essay detection. In this section, we describe the corpora used in the reviewed papers.

Higgins, Burstein and Attali (2006) evaluated two datasets, one for each type of off-topic essay. For the unexpected topic type, a dataset with 8.000 essays from students with level ranging from 6^{th} to 12^{th} grade was used. For the bad-faith type, 3.138 essays were selected from GMAT, GRE and TOEFL high-stakes tests and manually labeled.

Louis and Higgins (2010) used four different corpora of essays collected from TOEFL and GRE tests, including both learners and advanced English writers. In the study, the authors randomly sampled 350 essays on 7 prompts for the evaluation set, and used essays from 3 prompts as development data. The results of the novel approach were presented only for two of the four corpora, in which the prompts had very short statements (9-13 words in average).

Li and Yan (2012) evaluated 2.041 essays submitted to a large-scale test of English as a second language in China, known as CET4. In order to make the study feasible, substantial effort was spent in transcribing the originally handwritten essays to an electronic format. Therefore, the authors draw attention to the need to use the computer instead of handwriting in the essay test application.

Persing and Ng (2014) extracted 830 argumentative essays from the International Corpus of Learner English (ICLE), a publicly available corpus which consists of "more than 6.000 essays written by university undergraduates from 16 countries and 16 natives languages who are learners of English as a Foreign Language" [Persing and Ng 2014]. The authors asked human annotators to score the essays regarding the prompt adherence criterion within the range [1, 4]. By analyzing the

doubly annotated scores of 707 essays, the authors found a weak correlation (*r* .243).

Chen and Zhang (2016) had the largest dataset among the reviewed papers, which consists of four corpora containing about 200.000 essays each. These essays were randomly selected from two large-scale high-stakes tests: a college level and an English proficiency test. The authors selected 380 to 24.244 off-topic essays and the same number of on-topic essays from each of the four corpora to build an evaluation dataset of 57.176 essays. It is important to mention that all the reviewed studies used corpora of essays written only in English. Thus, this research field lacks corpora for other languages such as Portuguese.

4.4. Performance measures (Q3)

The third research question is about the performance results in the existing literature on automatic off-topic essay detection. The performance measured in the reviewed studies is presented in Table 9. Li and Yan (2012) have evaluated their results by visually analyzing the relation of the predicted prompt-adherence value and the overall essays score. However, the authors did not use a numeric measurement of accuracy, so their study is not mentioned in Table 9.

Several performance measures have been applied in the existing literature to evaluate off-topic essay detection systems. It is worth noting that the % FP (rate of false positives), % FN (rate of false negatives), precision, recall, F-score and % wrong prediction (WP) measures are similar since they all can be extracted from a confusion matrix. However, these measures cannot be directly compared and each has specific applications, being more or less suitable for different scenarios.

Table 9. Performance measured in existing literature

Reference	Approach	Results
[Higgins, Burstein and Attali 2006]	Model A	FP: 5.0% | FN: 30-38.0%
	Model B	FP: 4.7% | FN: 16.8-28.2%
	Model C_{UT}	FP: 6.8% | FN: 22.9%
	Model C_{BF}	FP: 3.0% | FN: 25.7%
[Louis and Higgins 2010]	Best	FP: 1.47-9.02% | FN: 9,02-11.97%
[Persing and Ng 2014]	Best	WP: .488 | MAE: .348 | MSE: .197 | *r*: .360
[Chen and Zhang 2016]	Best	Prec.: 100% | Recall: 2.2-18.1% | F-score: 4.4-30.7%

As shown in Table 9, in the studies where prompt-adherence was addressed as a binary classification task (on-topic or off-topic), the recall and false negatives rates varied in the range 2.2-38%. This means that the proposed approaches still can't detect a considerable amount of off-topic essays samples which were present in the evaluated dataset. Similarly, in the study carried out by Persing and Ng (2014), where regression models were built to predict a continuous value, a very high wrong predictions rate was found (48.8%) and a weak to moderate linear correlation (.360).

Off-topic essays may be very different: while some may present too much information copied from the prompt, others may have not nearly addressed the expected topic. With this, the several techniques presented in the literature may be more suitable for one type of off-topic essay than for another. The analysis of the results on each type

of problem can improve the model evaluation; however, it demands a labeled dataset.

5. Conclusion

This paper presents a systematic review on automatic off-topic essay detection. We reviewed the techniques and resources, the corpora and the performance measured in the existing literature.

The results found indicate some gaps and deficiencies in the existing approaches for off-topic essay detection. Among them, it can be mentioned: (i) current tested approaches have shown high error rates; (ii) existing studies have mostly used artificial essays sets for validation, which may have caused results inconsistent with real scenarios – for instance, in [Chen and Zhang 2016] a very low rate of recall was found while evaluating a real set of off-topic essays; (iii) despite the importance of analyzing results with respect to more than one off-topic essay type (e.g. unexpected topic and bad-faith), the approaches presented in the reviewed literature have mostly treated off-topic essays as a single category, and (iv) none of the reviewed studies have been applied to essays in the Portuguese language.

References

Chen, J. and 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, v. 140, August, p. 315–326.

Higgins, D., Burstein, J. and Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. Natural Language Engineering, v. 12, n. 2.

Higgins, D. and Heilman, M. (2014). Managing what we can measure: Quantifying the susceptibility of automated scoring systems to gaming behavior. Educational Measurement: Issues and Practice, v. 33, n. 3, p. 36–46.

Kitchenham, B. and Charters, S. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. EBSE Technical Report.

Li, Y. and 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, p. 65–68.

Louis, A. and 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, p. 92–95.

Persing, I. and Ng, V. (2014). Modeling Prompt Adherence in Student Essays. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, June, p. 1534–1543.

Santos, J. J., Paiva, R. and Bittencourt, I. I. (2016). Lexical-Syntactic Evaluation of written activities based on Genetic Algorithm and Natural Language Processing: An experiment on ENEM (in Portuguese). Brazilian Journal of Computers in Education, v. 24, n. 2, p. 92–107.

Silva, A. C., Barbosa, W., Rodrigues, C. L. and Ferreira, D. (2016). Technologies for teaching DHI students using Libras: a systematic review. Brazilian Symposium on Computers in Education, v. 27, n. 1, p. 747–756.