Natural Language Engineering **12** (2): 145–159. © 2006 Cambridge University Press
doi:10.1017/S1351324906004189 Printed in the United Kingdom

Identifying off-topic student essays without topic-specific training data†

D. HIGGINS, J. BURSTEIN and Y. ATTALI
Educational Testing Service, Rosedale Road,
Princeton, NJ 08541, USA
e-mail: dhiggins@ets.org

(*Received 1 May 2005; revised 1 November 2005*)

Abstract

Educational assessment applications, as well as other natural-language interfaces, need some mechanism for validating user responses. If the input provided to the system is infelicitous or uncooperative, the proper response may be to simply reject it, to route it to a bin for special processing, or to ask the user to modify the input. If problematic user input is instead handled as if it were the system's normal input, this may degrade users' confidence in the software, or suggest ways in which they might try to "game" the system. Our specific task in this domain is the identification of student essays which are "off-topic", or not written to the test question topic. Identification of off-topic essays is of great importance for the commercial essay evaluation system *Criterion*^{SM}. The previous methods used for this task required 200–300 human scored essays for training purposes. However, there are situations in which no essays are available for training, such as when users (teachers) wish to spontaneously write a new topic for their students. For these kinds of cases, we need a system that works reliably without training data. This paper describes an algorithm that detects when a student's essay is off-topic without requiring a set of topic-specific essays for training. This new system is comparable in performance to previous models which require topic-specific essays for training, and provides more detailed information about the way in which an essay diverges from the requested essay topic.

1 Introduction

Research problems in text document classification include automatic cataloguing of news articles (Allan, Carbonell, Doddington, Yamron and Yang 1998; Billsus and Pazzani 1999), sorting of e-mail (Sahami, Dumais, Heckerman and Horvitz 1998; Cohen, Carvalho and Mitchell 2004), internet-based search engines (McCallum, Nigram, Rennie and Seymore 1999; Joachims 2002), and classifying information in medical reports (Hripcsak, Friedman, Alderson, DuMouchel, Johnson and Clayton 1995; Wilcox and Hripcsak 2003; Chapman, Christensen, Wagner, Haug, Ivanov, Dowling and Olszewski 2005). Our research problem also relates to text classification,

† The authors would like to thank Chi Lu and Slava Andreyev for their help in carrying out the experiments described in this paper.

https://doi.org/10.1017/S1351324906004189 Published online by Cambridge University Press

but in an educational domain: automated essay evaluation. There is a great deal of previous work in automated essay scoring (Page 1966; Burstein, Kukich, Wolff, Lu, Chodorow, Braden-Harder and Harris 1998; Foltz, Kintsch and Landauer 1998; Larkey 1998; Elliot 2003), but our problem is a bit different. Specifically, our task is to evaluate if a student has written an *off-topic* essay (Burstein 2003).

The development of an off-topic essay detection capability is intended to enhance *Criterion*℠, a web-based, commercial essay evaluation system for writing instruction (Burstein *et al.* 2004). *Criterion* contains two complementary applications, for scoring and writing-analysis components. The scoring application, *e-rater*®, extracts linguistically-based features from an essay and uses a statistical model of how these features are related to overall writing quality to assign a score to the essay, typically on a scoring scale of 1 (worst) to 6 (best). The writing analysis application, *Critique*, is comprised of a suite of programs that evaluate errors in grammar, usage, and mechanics, identify an essay's discourse structure, and recognize undesirable stylistic features.

Though neither *e-rater* nor *Critique* contain components to evaluate off-topic writing, *Criterion* does provide such feedback to students. For training purposes, however, the method currently deployed in *Criterion* requires a significant number (200–300) of human-reader scored essays that are written to a particular test question (topic). This can be problematic, since *Criterion* allows users (teachers) to spontaneously write new topics for their students. (Essay *topics*, or *prompts*, are the short text passages which tell students what subject to write about, and how to structure their essays. They typically range from about 20 to 250 words in length.) In addition, *Criterion* content developers may add new topics to the system periodically. In neither case is there a chance to collect and manually score the requisite number of essay responses to train a new model. Another weakness of the current method is that it addresses different kinds of off-topic writing (discussed below) in the same way.

In this paper, we describe a method for identifying off-topic essays that does not require a large set of topic-specific training data. In addition, we provide a more fully-articulated model of off-topic essay detection which captures two different kinds of off-topic writing: *unexpected topic* essays and *bad-faith* essays. The differences between these two are described in Section 2.

2 What do we mean by off-topic?

Criterion's current model of off-topic essays assumes that they are all basically the same, or at least that they differ from on-topic essays in the same way. In fact, though, examination of the data reveals that off-topic essays are much like Tolstoy's "unhappy families".[1] On-topic essays tend to be very similar, but off-topic essays can be off-topic in quite divergent ways.

[1] *Anna Karenina*, first lines: "Happy families are all alike; every unhappy family is unhappy in its own way."

https://doi.org/10.1017/S1351324906004189 Published online by Cambridge University Press

```
Dear Computer!
I know that you are not going to let me continue this test if
I don‘t write my opinion to your argument here. But please be
aware that I‘m a very spesial student (by the way also a
smart girl). Fortunately I don’t have to do this section
according to the test regulation of my school. I know, that
with this small peace of unimportant information your
software programm will jump now stright away to the next test
section, that I actually also don’t need to write . I
appriciate very mach your patiance and hope that you will not
be dissappointed, or even offended by my lazziness.
                With best regards sincerely yours
                Maria
```

Fig. 1. Sample bad-faith essay.

We can identify the following broad classes of off-topic essays:

The empty essay is one which is so ridiculously short that it cannot be processed by our essay analysis engine.

The unexpected-topic essay is a well-formed, possibly very well-written essay, on a topic that does not respond to the expected test question. This can happen if a student inadvertently cuts-and-pastes the wrong essay that s/he has prepared off-line.

The banging-on-the-keyboard essay is dominated by gibberish text such as "`alfjdla dfadjflk ddjdj8ujdn`," with little real lexical content.

The copied-prompt essay is one which consists entirely or primarily of text copied and pasted from the essay topic itself. The prompt may be reproduced in its entirety, or only selected sentences may be included.

The irrelevant musings essay is one in which the examinee enters a chunk of text which is of substantial length, and may even be fairly coherent, but which is entirely unrelated to the essay topic on which they were asked to write. An example is shown in Figure 1. This is the most common sort of off-topic essay in practice, as many students respond uncooperatively out of boredom or hostility to the teacher.

Of course, a given "weird" essay may also be a combination of these types, or may not fit into this scheme at all. This list is meant to be an enumeration of the off-topic essay types commonly encountered in our experience with *Criterion*, rather than an exhaustive list.

Any of these cases may also happen when users just want to try to fool the system. *Criterion* users are concerned if any sort of off-topic essay fails to be recognized as such by the system. Even if a low essay score would be assigned, we clearly want to prevent the system from scoring such "essays" at all. *Empty* essays are handled by a filter which flags essays which do not meet the minimum length requirements for *Criterion*. *Banging-on-the-keyboard* essays are handled by a capability in *Criterion* that considers ill-formed part-of-speech sequences in an essay.

https://doi.org/10.1017/S1351324906004189 Published online by Cambridge University Press

The other types of off-topic essays are left for the models described in this paper. In the following sections, we will treat *copied-prompt* and *irrelevant musings* essays together under the title *bad-faith* essays, because they have in common the fact that they are the result of examinees' deliberate non-responsiveness to the essay task. These two off-topic essay types are also frequently combined to some degree, and are largely amenable to the same sort of predictive model.

3 Methods of off-topic essay identification

3.1 Data

Before describing the methods we use to determine whether an essay is on-topic or off-topic, we introduce here the data sets on which these methods will be evaluated in the remainder of the paper. Two sets of data are used for these experiments: the *unexpected topic* essay set and *bad-faith* essay set, corresponding to the distinction made above.

The data that we used to evaluate the detection of *unexpected topic* essays contain a total of 8,000 student essays. Within these 8,000 are essays written to 36 different test questions (i.e., prompts or topics), approximately 225 essays per topic. The level of essay spans from the 6^{th} through 12^{th} grade. There is an average of five topics per grade. These data are all good faith essays that were written to the expected topic. Note, however, that on-topic essays for one prompt can be used as exemplars of unexpected-topic essays for another prompt in evaluating our systems.

The data used to evaluate the detection of *bad-faith* essays consist of three separate sets of essays: one set drawn from GMAT® test essays, one set drawn from GRE® test essays, and one set drawn from TOEFL® test essays. Ideally, we would like to have been able to evaluate *bad-faith* essay detection on essays drawn from the target population of elementary and secondary-school students. However, data for these college programs was much easier to obtain in large quantities. There will certainly be some differences between these sources of data, but we assume that off-topic essays differ from on-topic ones in basically the same ways, regardless of students' age.

Each set of *bad-faith* essays was drawn from among a set of essays assigned a score of 0 by human graders. We then excluded essays from this set which did not meet the length requirement of containing at least two sentences, according to a simple tokenization algorithm. Finally, a human annotator read through each set of essays to ensure that they were indeed *bad-faith* essays, and not *unexpected topic* essays, or simply essays assigned a zero score for another reason. This procedure resulted in a set of 1288 GMAT essays, 539 GRE essays, and 1311 TOEFL essays.

None of the questions these essays were supposed to respond to were the same as the 36 test questions in the 6^{th} to 12^{th} grade pool of essay questions used for evaluating *unexpected topic* essay detection.

3.2 Content Vector Analysis

This section describes Content Vector Analysis (CVA), a vector-based semantic similarity measure which is used by two of the off-topic essay identification models

https://doi.org/10.1017/S1351324906004189 Published online by Cambridge University Press

introduced below. Readers familiar with CVA may wish to skip this section, referring back to it only for the specific details of the CVA models used below.

CVA is an information-retrieval method for quantifying the degree to which two texts share the same vocabulary. It involves constructing a content vector for the each of the two texts to be compared, in which each component of the vector corresponds to a certain word's frequency of occurrence in the text. The similarity between the texts is calculated as the cosine of the angle between these content vectors (Salton 1989). Basically, texts are gauged to be similar to the extent that they contain the same words in the same proportion.

We do not do any stemming to preprocess the texts for CVA in the models discussed below, but we do use a stoplist to exclude non-content-bearing words from the calculation. We use a variant of the `tf*idf` weighting scheme to associate weights with each word in a text's content vector. Specifically, the weight is given as $(1 + \log(tf)) \times \log(\frac{D}{df})$, where tf is the "term frequency", df is the "document frequency", and D is the total number of documents in the collection.

The term frequencies in this scheme are taken from the counts of each word in the document itself, of course (the essay or prompt text). The document frequencies in the models to be introduced are taken from various sources, however.

Our topic-specific models for off-topic essay identification draw document frequency statistics from the essay pool itself. This is the same term weighting scheme used by the *e-rater* scoring engine (Burstein *et al.* 2004), and is motivated by the idea that statistics gathered from documents of the same genre, addressing the same essay topic, will best reflect the relative importance of terms in the essays.

In our models which do not make use of topic-specific training data, though, we use document frequencies derived from the TIPSTER collection (Harman 1992), making the assumption that these document frequency statistics will be relatively stable across genres. The large size of this document collection helps to ensure the stability of the document frequency estimates we extract.

3.3 Two topic-specific methods of off-topic essay identification

We have explored two different methods of identifying off-topic essays which use training data specific to a given essay topic. In the following, we refer to our deployed model as **Model A**, and the newer model under development as **Model B**.

In the context of our topic-specific models, we do not make a distinction between *unexpected topic* and *bad-faith* essays. Both are considered together in the production of a single model to detect off-topic essays.

3.3.1 Model A

In our currently operational method of off-topic essay detection, we compute two values derived from a content vector analysis program used in *e-rater* for determining vocabulary usage in an essay (Burstein *et al.* 2004; Allan *et al.* 1998).[2]

[2] This method was developed and implemented by Martin Chodorow and Chi Lu.

https://doi.org/10.1017/S1351324906004189 Published online by Cambridge University Press

For each essay, *z-scores* are calculated for two variables:

1. relationship to words in a set of training essays written to a prompt (essay question), and
2. relationship to words in the text of the prompt.

The *z-score* value indicates a novel essay's relationship to the mean and standard deviation values of a particular variable based on a training corpus of human-scored data. The score range is usually 1 through 6, where 1 indicates a poorly written essay, and 6 indicates a well-written essay. To calculate a *z-score*, the mean value and the corresponding standard deviation (SD) for *maximum cosine* or *prompt cosine* are computed based on the human-scored training essays for a particular test question. The formula for calculating the z-score for an new novel essay is: $\text{z-score} = \frac{(\text{value}-\text{mean})}{\text{SD}}$ As this formula shows, the *z-score* indicates how many standard deviations from the mean our essay is on the selected dimension. For our task, z-scores are computed for:

1. the maximum cosine, which is the highest cosine value among all cosines between an unseen essay and all human-scored training essays, and
2. the prompt cosine, which is the cosine value between an essay and the text of the prompt (test question).

When a *z-score* exceeds a set threshold, it suggests that the essay is anomalous, since the threshold value indicates an acceptable distance from the mean.

We evaluate the accuracy of these approaches based on the false positive and false negative rates. The *false positive rate* is the percentage of appropriately written, on-topic essays that have been incorrectly identified as off-topic; the *false negative rate* is the percentage of true off-topic essays not identified (missed) as off-topic. Within a deployed system, it is preferable to have a lower false positive rate. That is, we are more concerned about telling a student, incorrectly, that s/he has written an off-topic essay, than we are about missing an off-topic essay.

For the *unexpected topic* essay set, the rate of false positives using this method is 5.0%, and the rate of false negatives is 38.0%, when the z-scores of both the *maximum cosine* and *prompt cosine* measures exceed the thresholds. For the *bad-faith* essay data, the false negative rates were 40.1% on the TOEFL set, 19.3% on the GRE set, and 24.6% on the GMAT set. Because this model does not distinguish between the different types of off-topic essays, we cannot further break down the false positive rate according to the kind of anomaly detected. It is not surprising that *bad-faith* essays should be hardest to identify in the TOEFL data set, since such essays tend to be of lower quality than GRE or GMAT essays.

3.3.2 Model B

A newer prompt-specific method has been developed recently that yields better performance. It is based on calculating two rates for each word used in essays:

1. the proportion of word occurrences across many topics (generic, or prompt-independent, rate), and
2. the proportion of word occurrences within a topic (prompt-specific rate).

https://doi.org/10.1017/S1351324906004189 Published online by Cambridge University Press

The generic rate of occurrence for each word (G_i) across the large sample is calculated one time only from a large sample of essays across different prompts from within one program, or within similar grade-levels. It is interpreted as the base-rate level of popularity of each word. The prompt-specific rate (S_i) is computed from a training sample of essays that were written to the specific prompt for which an individual essay is to be compared. These two rates are used to compute an overall index for each individual essay:

$$\frac{1}{N}\sum_{i=1}^{n}\sqrt{S_i(1-G_i)} \tag{1}$$

Equivalently, to compute this index, we carry out the following steps:

1. Identify S_i and G_i values for all words in an essay based on pre-determined values from training sets.
2. For each word, compute $S_i(1-G_i)$ and take the square root. Now sum these square roots over all words.
3. Multiply the sum of square roots by $\frac{1}{N}$, where N is the number of words in the essay, and the two rates are computed for all words in the essay.

A word not in either one of the training samples will have a rate of 0, so a totally new word, not in either the generic or the specific sample, will also have a weight of zero ($0 \times (1-0) = 0$). This index can be interpreted as an average of word weights, where the weights are larger for words that appear more frequently in the prompt-specific essays, but at the same time are not frequent in other prompts. These are the words that would most contribute to the discrimination between on-topic essays and off-topic essays. The range of word weights is from 0 (when a word never appears in the specific sample and/or always appears in the generic sample of essays) to 1 (when it appears in every specific essay but never appears in the generic sample). The use of the square-root transformation in the weighting of words is designed to emphasize heavily-weighted words over low-weighted words and slightly improves performance.

The classification of new essays as off- or on-topic is determined by setting a cutoff on the index values. This cutoff is based on the distribution of index values in the prompt-specific training sample. For example, the cutoff could be set to be equal to the fifth percentile value in this distribution. (Setting the cutoff in this way would fix the false positive rate at 5% for the training sample.)

Two differences between Model B and Model A are worth noting. First, the new weighting system weights more heavily frequent words which appear in many essays written to a particular prompt, than words that appear infrequently in essays, contrary to the z-score-based system. Second, Model A does not distinguish between words that are frequent in every essay regardless of topic, like *want* or *things* and words that are much more frequent in certain topics, like *president*. Even words like *school* or *work*, which are clearly topic-dependent, are nonetheless very frequent in general because many K-12 essay prompts are related to these issues. As a consequence, their weight should still be discounted even in prompts that are related

https://doi.org/10.1017/S1351324906004189 Published online by Cambridge University Press

to these issues, because the occurrence of the words in an essay is not a good discriminative sign for off-topicness.

For this newer method, the rate of false positives is 4.7%, and the rate of false negatives is 28.2%. For the *bad-faith* essay data, the false negative rates were 28.9% on the TOEFL set, 8.7% on the GRE set, and 7.8% on the GMAT set. Again, this model does not distinguish between different sorts of off topic essays, so the false positive rate can only be given as an overall number. Unfortunately, this new and improved method still requires the topic-specific sets of human-scored essays for training.

3.4 Identifying off-topic essays using CVA and no topic-specific training data

Most recently, we have developed a model for assessing whether an essay is on-topic or off-topic without topic-specific training data. In addition, this model handles *unexpected topic* and *bad-faith* essay detection independently, providing more information about the reason an essay is off-topic. It makes sense to treat these two different types of off-topic essays separately, because of their very different properties. *Unexpected topic* essays differ from on-topic essays only in their subject matter, whereas *bad-faith* have a number of other characteristics which distinguish them from normal, on-topic essays, including length, vocabulary usage, and the frequency of unrecognized words. We will refer to the topic-independent model for identifying *unexpected topic* essays as Model C_{UT}, and to the corresponding model for identifying *bad-faith* essays as Model C_{BF}.

3.4.1 Unexpected topic essays (Model C_{UT})

Our topic-independent model for off-topic essay detection uses CVA[3], and also relies on similarity scores computed between new essays and the text of the prompt on which the essay is supposed to have been written. Unlike Models A and B, this method does not rely on a pre-specified similarity score cutoff to determine whether an essay is on- or off-topic. Because this method is not dependent on a similarity cutoff, it also does not require any prompt-specific essay data for training in order to set the value of this parameter.

Instead of using a similarity cutoff, our newer method uses a set of *reference essay prompts*, to which a new essay is compared. The similarity scores from all of the essay-prompt comparisons, including the similarity score that is generated by comparing the essay to the target prompt, are calculated and sorted. If the target prompt is ranked amongst the top few vis-à-vis its similarity score, then the essay is considered on topic. Otherwise, it is identified as off topic.

[3] We experimented with another vector-based similarity measure, namely Random Indexing (RI) (Sahlgren 2001), and CVA had better performance. The tendency of RI, LSA, and other reduced-dimensionality vector-based approaches to assign higher similarity scores to texts that contain similar (but not the same) vocabulary may be a contributing factor. The fact that an essay contains the exact words used in the prompt is an important clue that it is on topic, and this may be obscured using an approach like RI.

https://doi.org/10.1017/S1351324906004189 Published online by Cambridge University Press

This new method utilizes information that is available within *Criterion*, and does not require any additional data collection of student essays or test questions.

We know from previous experimentation that essays tend to have a significant amount of vocabulary overlap, even across topics, as do the test questions themselves. For instance, if one topic is about "school" and another topic is about "teachers," essays written to these topics are likely to use similar vocabulary. Even more generally, there is a sublanguage of essays that may be referred to as generic word use. In the sublanguage of standardized test essays are words such as "I," "agree," and "opinion." Therefore, selecting a threshold based on any measure to estimate similar vocabulary usage between an essay and the essay question has proven to be ineffective. Specifically, the similarity of essays to their (correct) prompt can be highly variable, which makes it impossible to set an absolute similarity cutoff to determine if an essay is on an unexpected topic. However, we can be fairly certain that the target prompt should at least rank among the *most* similar, if the essay is indeed on topic. Given this, we carried out the evaluation in the following way.

Starting with our 36 prompts (topics), we performed an 18-fold cross-validation. For each fold, we use 34 *reference prompts*, and two *test prompts*. This cross-validation setup allows us to distinguish two different evaluation conditions. The first, *training set performance*, is the system's accuracy in classifying essays that were written on one of the reference prompts. The second, *test set performance*, is the accuracy of the system in classifying essays written on one of the test prompts.

For each cross-validation fold, each essay from across the 34 reference prompts is compared to the 34 reference prompt texts, using the cosine correlation value from CVA.[4] Therefore, an essay is compared to the actual prompt to which it was written, and an additional 33 prompts on a different, unexpected topic. Based on the computed essay-prompt cosine correlation value, essays are considered 'on-topic' only if the value is among the top N values; otherwise the essay is considered to be off-topic. So, for instance, if the similarity value is amongst the top 5 of 34 values (top 15%), then the essay is considered to be on-topic. This gives rise to the training set performance shown in Figure 2. The essays written to the test prompts are also evaluated. If A and B are the two test prompts, then all essays on prompt A are compared to the 34 reference prompts plus prompt A, while all essays on prompt B are compared to the 34 reference prompts plus prompt B. The resulting rankings of the prompts by similarity are used to determine whether each test essay is correctly identified as on-topic, producing the false positive rates for the training set in Figure 2. Finally, all essays on prompt A are compared to the 34 reference prompts plus prompt B, while all essays on prompt B are compared to the 34 reference prompts plus prompt A. This allows us to generate the false negative rates for the training set in Figure 2.

[4] We also experimented with using only the first N content words of each essay in the CVA comparison, with the idea that the topic of the essay was likely to be established early on. However, this did not provide better results than using the entire essay in the similarity calculation, so this approach was abandoned. We speculate that this is because in a longer span of text, the vocabulary used is less likely to be misleading about the essay topic.

https://doi.org/10.1017/S1351324906004189 Published online by Cambridge University Press

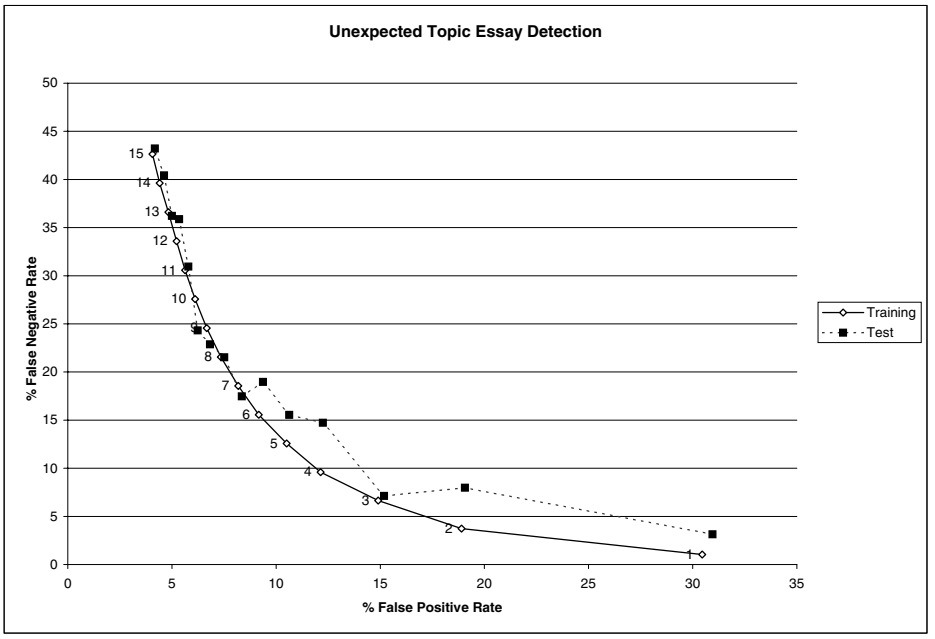


Fig. 2. Performance of CVA-based model in predicting unexpected-topic essays.

Figure 2 shows the trade-off between the false positive rate and the false negative rate in our model of unexpected-topic essay detection. The number labeling each point on the graph indicates the cutoff N, which is the number of prompts considered close enough to the essay to be regarded as on-topic. We see in Figure 2 that the best choice of this parameter for our application is probably around 10, which gives us a false positive rate of 6.8% and a false negative rate of 22.9% on test data. These rates are achieved without the use of labeled training data, and represent only a moderate degradation in performance compared to Model B.

3.4.2 Bad-faith essays (Model C_{BF})

For identifying *bad-faith* essays, a different sort of model is called for, because we do not expect these essays to share much vocabulary with any essay prompt. These are the worst-case off-topic essays, where no attempt was made to answer any kind of essay question.

As discussed above, there are a number of features besides similarity of vocabulary use which serve to identify essays which have been written in *bad faith*. The most telling feature is the length of the essay, since the *irrelevant musings* (cf. section 2) type of essay tends to be quite short. One might also think that long stretches of text which are also found in the prompt would be a good feature, since we are also interested in the *copied prompt* essay type here. However, *copied prompt* essays do not always contain perfectly-rendered text from the prompt, as would result from simply using a text editor's "copy and paste" functions. Sometimes users type in text

https://doi.org/10.1017/S1351324906004189 Published online by Cambridge University Press

from the essay prompt with spelling errors, omissions, or added elements, so our features have to be a bit fuzzier to account for this.

The following five features of an essay are used by Model C_{BF} to predict whether the essay was written in *bad faith*:

1. The CVA similarity between the essay text and the essay prompt
2. The number of words in the essay (excluding stopwords)
3. The proportion of words (excluding stopwords) which are found in the essay, but not in the prompt
4. The ratio of word types to word tokens in the essay
5. The frequency with which markers of 'direct address' occur in the essay

This last feature is motivated by the fact that *bad-faith* essays often contain statements directed to someone (such as the test author or teacher), rather than just the sort of expository writing one would expect to find in an essay. Examples of these sorts of markers, which suggest that the student may be writing *about* the test *to* someone, include the name of the test (e.g., "GMAT"), the word "hello", and the word "thanks".

We use a support vector machine (SVM) to make the prediction whether an essay should be flagged as *bad-faith* or not based on these features. SVMs are well-suited to this application because they allow for automatic non-linear combination of real-valued input features, and show good generalization with limited training data (Vapnik 1995; Christianini and Shawe-Taylor 2000). We use an SVM with a radial basis function kernel and a cost parameter of 100.

For each of our human-verified *bad-faith* essays, we constructed a vector of the above-listed features, derived from the essay text and the prompt to which it was supposed to have been written. We then added it to the data set for this experiment as a positive exemplar of a *bad-faith* essay. In a similar fashion, we used the essays from the *unexpected topic* essay set to produce the negative exemplars for this experiment. Pairing each essay with the topic on which it was actually written provides us with data on what on-topic essays look like, and how they are related to their prompts, so that Model C_{BF} can distinguish these from *bad-faith* essays.

This set of positive and negative exemplars was then randomly divided into ten subsets for ten-fold cross-validation. On each training run, $\frac{9}{10}$ of the data was used to train the SVM, and the other tenth was used as a test set. Figure 3 shows the results on the test sets, aggregated across all ten cross-validation runs. Results for the GMAT, GRE, and TOEFL *bad-faith* essay sets are reported separately, and for each set, a range of results is reported to show the tradeoff between the false negative rate and the false positive rate. These different weightings of false negatives and false positives were achieved by differentially weighting the positive and negative exemplars during SVM training.

Once again the bad-faith essays of the TOEFL data set are the most difficult to identify, and the performance on the GMAT data set is the best. Reasonable parameterizations for our purposes would yield a false positive rates of 5.9% on TOEFL, 0.90% on GRE, and 0.94% on GMAT, with false negative rates of 36.8% on TOEFL, 30.0% on GRE, and 12.7% on GMAT.

https://doi.org/10.1017/S1351324906004189 Published online by Cambridge University Press

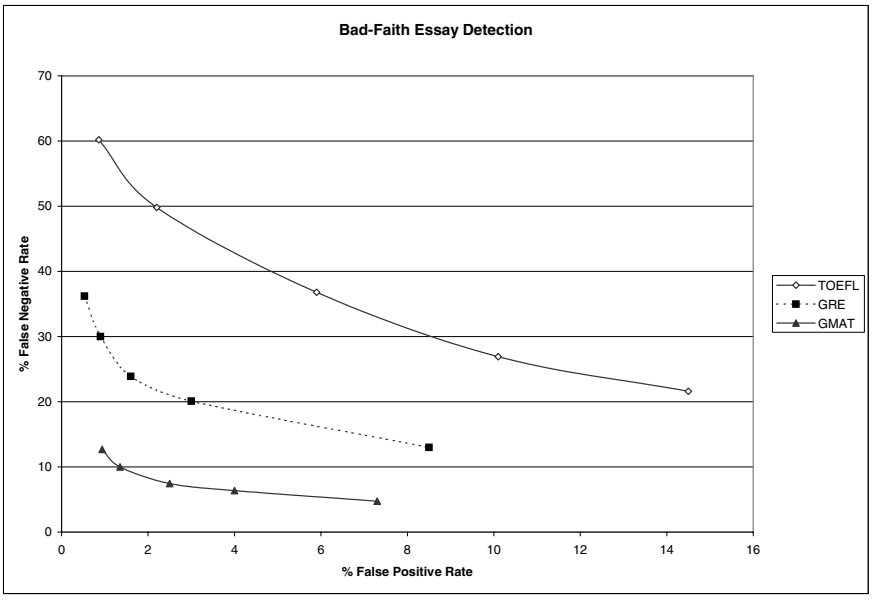


Fig. 3. Performance of CVA-based model in predicting bad-faith essays.

To see that the full SVM model we have introduced is really necessary to the task, we can compare a simple baseline model which uses only the CVA similarity between essay and prompt to determine whether an essay is *bad-faith* or not. Linear models of this sort perform very badly on all data sets, because essay-prompt similarity is correlated positively with *copied-prompt* essays and negatively with *irrelevant musings* essays. The best RBF-kernel-based models using only this feature yield false positive rates of 15.3% on TOEFL, 8.6% on GRE, and 13.2% on GMAT, and false negative rates of 39.0% on TOEFL, 18.0% on GRE, and 14.4% on GMAT.

4 Model comparison

To properly assess the models introduced here, we first need to be explicit about how each would be used in an application. Figure 4 indicates the intended process flow for each model. Models A and B simply make a distinction between on-topic and off-topic essays, so the process flow is very simple. If the model flags an essay as off-topic, it is handled specially; otherwise, the essay is sent on for scoring.

Model C is slightly more complex, though, because it consists of the submodels C_{BF} and C_{UT}, which respectively detect *bad-faith* and *unexpected topic* essays. Any essays not flagged by Model C_{BF} are sent on to Model C_{UT}, which makes the final determination whether the essay will be sent on for scoring.

Table 1 reprises the performance numbers for each model reported earlier in the paper, in a format which facilitates comparison. Where possible, the performance numbers are broken down according to the subtasks of *bad-faith* and *unexpected topic* essay detection, but we also provide a single "general off-topic" essay

https://doi.org/10.1017/S1351324906004189 Published online by Cambridge University Press

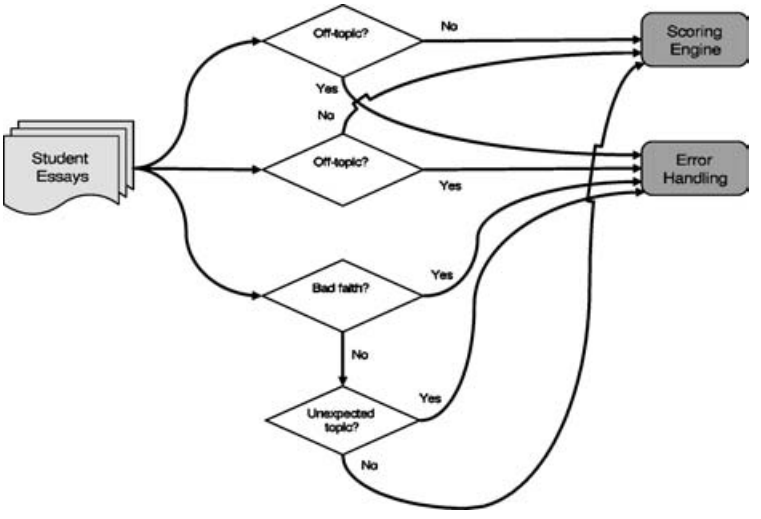


Fig. 4. System architecture for Models A, B, and C.

Table 1. *Comparison of Models A, B and C*

	Bad Faith Essay Detection		Unexpected Topic Essay Detection		General Off-Topic Essay Detection	
	FP%	FN%	FP%	FN%	FP%	FN%
Model A	–	30.0	–	38.0	5.0	30.0%–38.0%
Model B	–	16.8	–	28.2	4.7	16.8%–28.2%
Model C	3.0	25.7	6.8	22.9	6.8–9.6	22.9%–25.7%

identification category, which considers only the models' success in separating both types of off-topic essays from on-topic ones.

Because the structure of Model C differs from that of Models A and B, different sets of performance numbers must be given. We can report false negative rates for both *bad-faith* and *unexpected topic* essays for each model, by simply running the bad essays from each set through the models and seeing how many fail to be flagged. For Models A and B, however, we can only provide a single false positive rate, because these models do not flag *bad-faith* and *unexpected topic* essays differently. Finally, we cannot give a precise overall false negative rate, including both types of off-topic essays, for any of the three models. To do this, we would need an estimate of the relative frequency of *bad-faith* and *unexpected topic* essays, which we do not have. We do know, however, that the overall false negative rate will fall somewhere between the false negative rates for *bad-faith unexpected topic* essays.

The performance numbers for *bad-faith* essay detection in Table 1 are an average across the TOEFL, GRE, and GMAT essays in the *bad-faith* data set, weighted by the number of essays in each set. This is probably a reasonable estimate of the

https://doi.org/10.1017/S1351324906004189 Published online by Cambridge University Press

system's performance on essays from *Criterion*, since we expect these essays to be higher-quality than TOEFL essays, but not as clean as those from GRE or GMAT.

The overall false positive rate for Model C (the percentage of on-topic essays which are labeled either as *bad-faith* or as *unexpected topic* essays) depends on the degree of independence between Models C_{BF} and C_{UT}. In the worst case, there would be no overlap between the sets of essays which the two submodels incorrectly identify as off-topic. In that case the overall false positive rate would be $0.03 + (1 - 0.03) \times 0.068 = 0.096$. In the best case, the misidentifications of the two models would coincide, and the false positive rate would be $\max(0.03, 0.068) = 0.068$. We suspect that the actual false positive rate will be closer to the lower end of the range, due to the similar statistical information adduced by the two submodels.

Table 1 shows that the topic-specific methods, in particular Model B, outperform Model C in our most critical aspect, the overall false positive rate. When topic-specific training data is available, they are clearly preferable. However, when such data is not available, or when further information is needed about whether an essay is of the *bad-faith* or *unexpected topic* type, Model C's performance is good enough that it will be useful in many assessment contexts. Model C's comparatively higher false positive rate on the task of identifying *unexpected topic* essays could perhaps be offset by routing the essays flagged by the system to a less drastic processing step, such as asking the user to confirm that they have entered the proper essay, rather than rejecting it outright.

5 Discussion and conclusions

Criterion^{SM}is an on-line essay evaluation service with over 500,000 subscribers. Currently, the system has only a supervised algorithm for detecting off-topic essays input by student writers. Since this method requires 200–300 human-scored essays to train each new essay question, the application cannot provide feedback about off-topic writing for topics entered on-the-fly by instructors. By the same token, if *Criterion* content developers want to periodically add new essay questions, off-topic essay detection cannot be applied until sufficient human-scored data are collected. In addition, the current supervised method treats all off-topic essays alike.

In this study, we have developed an unsupervised algorithm that requires only text of existing essay questions, the text of the new essay question, and the student essay in order to predict off-topicness. Our method also makes a distinction between two kinds of off-topic essays: *unexpected topic* and *bad-faith* essays. While topic-specific models remain superior on the narrow task of distinguishing on-topic essays from off-topic ones, our new topic-independent models have sufficiently good performance for use when topic-specific models are not available, or when more detailed information is needed about the way in which a given essay fails to address the topic.

References

Allan, J. Carbonell, J., Doddington, G., Yamron, J. and Yang, Y. (1998) Topic detection and tracking pilot study: final report. *Proceedings of the Broadcast News Transcription and Understanding Workshop*, pp. 194–218.

https://doi.org/10.1017/S1351324906004189 Published online by Cambridge University Press

Attali, Y. and Burstein, J. (2004) Automated essay scoring with *e-rater* V.2.0. Presentation at the Annual Meeting of the International Association for Educational Assessment, Philadelphia, PA.

Billsus, D. and Pazzani, M. (1999) A hybrid user model for news story classification. *Proceedings of the Seventh International Conference on User Modeling (UM '99)*.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L. and Harris, M. (1998) Automated scoring using a hybrid feature identification technique. *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics*, pp. 206–210.

Burstein, J., Chodorow, M. and Leacock, C. (2004) Automated essay evaluation: The *Criterion* online writing service. *AI Magazine* **25**(3): 27–36.

Burstein, J. (2003) The *e-rater*® scoring engine: Automated essay scoring with natural language processing. In: M. Shermis and J. Burstein (eds.), *Automated Essay Scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum.

Chapman, W., Christensen, L., Wagner, M., Haug, P., Ivanov, O., Dowling, J. and Olszewski, R. (2005) Classifying free-text triage chief complaints into syndromic categories with natural language processing. *Artificial Intelligence in Medicine* **33**(1): 30–40.

Christianini, N. and Shawe-Taylor, J. (2000) *Support Vector Machines and other Kernel-based Learning Methods*. Cambridge, UK: Cambridge University Press.

Cohen, W., Carvalho V. and Mitchell, T. (2004) Learning to classify email into "Speech Acts". Proceedings of EMNLP 2004, pp. 309–316.

Elliott, S. (2003) Intellimetric: From here to validity. In: M. Shermis and J. Burstein (eds.), *Automated Essay Scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum.

Foltz, P., Kintsch, W. and Landauer, T. (1998) Analysis of Text Coherence Using Latent Semantic Analysis. *Discourse Processes* **25**(2–3): 285–307.

Harman, D. (1992) The DARPA TIPSTER project. *SIGIR Forum* **26**(2): 26–28.

Hripcsak, G., Friedman, C., Alderson, P., DuMouchel, W., Johnson, S. and Clayton, P. (1995) Unlocking clinical data from narrative reports: a study of natural language processing. *Annals of Internal Medicine* **122**(9): 681–688.

Joachims, T. (2002) Optimizing search engines using clickthrough data. *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*

Larkey, L. (1998) Automatic essay grading using text categorization techniques. *Proceedings of the 21st ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 90–95.

McCallum, A., Nigam, K., Rennie, J. and Seymore, K. (1999) Building Domain-Specific Search Engines with Machine Learning Techniques. *Proceedings of the AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*.

Page, E. (1966) The imminence of grading essays by computer. *Phi Delta Kappan* **48**: 238–243.

Sahami, M., Dumais, S., Heckerman, D. and Horvitz, E. (1998) A Bayesian approach to filtering junk email. *Learning for Text Categorization: Papers from the 1998 Workshop*. AAAI Technical Report WS-98-05.

Sahlgren, M. (2001) Vector-based semantic analysis: representing word meanings based on random labels. *Proceedings of the ESSLLI 2001 Workshop on Semantic Knowledge Acquisition and Categorisation*.

Salton, G. (1989) *Information Retrieval: Data Structures and Algorithms*. Reading, MA: Addison-Wesley.

Vapnik, V. (1995) *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Wilcox A. and Hripcsak G. (2003) The role of domain knowledge in automating medical text report classification. *Journal of the American Medical Informatics Association* **10**: 330–338.

https://doi.org/10.1017/S1351324906004189 Published online by Cambridge University Press