# Advanced Capabilities for Evaluating Student Writing: Detecting Off-Topic Essays Without Topic-Specific Training

Jill BURSTEIN
Derrick HIGGINS
*Educational Testing Service*
*Princeton, New Jersey, USA*

**Abstract.** We have developed a method to identify when a student essay is off-topic, i.e. the essay does not respond to the test question topic. This task is motivated by a real-world problem: detecting when students using a commercial essay evaluation system, *Criterion*[SM], enter off-topic essays. Sometimes this is done in bad faith to trick the system; other times it is inadvertent, and the student has cut-and-pasted the wrong selection into the system. All previous methods that perform this task require 200-300 human scored essays for training purposes. However, there are situations in which no essays are available for training, such as when a user (teacher) wants to spontaneously write a new topic for her 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. The system also distinguishes between two different kinds of off-topic writing. The results of our experiment indicate that the performance of this new system is comparable to the previous system that does require topic-specific essays for training, and conflates different types of off-topic writing.

## Introduction

Research problems in text document classification include sorting of e-mail ([17],[8]) internet-based search engines ([15],[13]), automatic cataloguing of news articles, ([1],[3]) and classifying information in medical reports ([12],[20],[7]). Our research problem also relates to text classification, but in an educational domain: automated essay evaluation. Much work has been done in this area with regard to automated essay scoring ([16], [4],[10],[14],[9]). Our problem is a bit different. Specifically, our task is to evaluate if a student has written an *off-topic* essay ([6]).

The context of this work is the development of an off-topic essay detection capability that will function within *Criterion*[SM], a web-based, commercial essay evaluation system for writing instruction ([5]). *Criterion* contains two complementary applications. 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 ranking (score) to the essay, typically on a scoring scale of 1 (worst) to 6 (best). The second application, *Critique*, is comprised of a suite of programs that evaluates errors in grammar, usage, and mechanics, identifies an essay's discourse structure, and recognizes undesirable stylistic features. *Criterion* currently has additional functionality that provides such feedback about off-topic writing to students. For training purposes, however, the current method requires a significant number (200-300) of human-reader scored essays that

are written to a particular test question (topic). This can be problematic in the following situation. *Criterion* allows users (teachers) to spontaneously write new topics for their students. In addition, *Criterion* content developers may also add new topics to the system periodically. In both cases, there is no chance to collect and manually score 200–300 essay responses. Another weakness of the current method is that it addresses different kinds of off-topic writing in the same way.

In this study, we have two central tasks: First, we want to develop a method for identifying off-topic essays that does not require a large set of topic-specific training data, and secondly, we also want to try to develop a method that captures two different kinds of off-topic writing: *unexpected topic essays* and *bad faith essays*. The differences between these two are described below.

In the remaining sections of this paper, we will define what we mean by an off-topic essay, discuss the current methods used for identifying off-topic essays, and introduce a new approach that uses content vector analysis, but does not require large sets of human-scored essay data for training. This new method can also distinguish between two kinds of off-topic essays.

## 1. What Do We Mean By Off-Topic?

Though there are a number of ways to form an off-topic essay, this paper will deal with only two types. In the first type, a student writes a well-formed, well-written essay on a topic that does not respond to the expected test question. We will refer to this as the *unexpected topic* essay. This can happen if a student inadvertently cuts-and-pastes the wrong essay that s/he has prepared off-line.

In another case, students enter a *bad faith* essay into the application, such as the following:

> "*You are stupid. You are stupid because you can't read. You are also stupid becuase you don't speak English and because you can't add.*
>
> *Your so stupid, you can't even add! Once, a teacher give you a very simple math problem; it was 1+1=?. Now keep in mind that this was in fourth grade, when you should have known the answer. You said it was 23! I laughed so hard I almost wet my pants! How much more stupid can you be?!*
>
> *So have I proved it? Don't you agree that your the stupidest person on earth? I mean, you can't read, speak English, or add. Let's face it, your a moron, no, an idiot, no, even worse, you're an imbosol.*"

Both cases may also happen when users just want to try to fool the system. And, *Criterion* users are concerned if either type is not recognized as off-topic by the system. A third kind of off-topic essay is what we call the *banging on the keyboard* essay, e.g., "*alfjdla dfadjflk ddjdj8ujdn.*" This kind of essay is handled by an existing capability in *Criterion* that considers ill-formed syntactic structures in an essay.[1] In the two cases that we consider, the essay is generally well-formed in terms of its structure, but it is written without regard to the test question topic. Another kind of off-topic writing could be a piece of writing that contains any combination of *unexpected topic*, *bad-faith*, or *banging on the keyboard* type texts. In this paper, we deal only with the *unexpected topic* and *bad-faith* essays.

## 2. Methods of Off-Topic Essay Identification

### 2.1 Computing Z-scores, Using Topic-Specific Essays for Training

---

[1] This method was developed by Thomas Morton.

In our current 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 ([5],[1]).[2] Off-topic in this context means that a new, unseen essay appears different from other essays in a training corpus, based on word usage, or, an essay does not have a strong relationship to the essay question text. Distinctions are not necessarily made between *unexpected topic* or a *bad faith* essays.

For each essay, *z-scores* are calculated for two variables: a) relationship to words in a set of training essays written to a prompt (essay question), and b) 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 essay 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.[3] For our task, z-*scores* are computed for: a) the *maximum cosine,* which is the highest cosine value among all cosines between an unseen essay and all human-scored training essays, and b) 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 typically indicates a value representing 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[4], the rate of false positives using this method is approximately 5%, and the rate of false negatives is 37%, when the z-scores of both the *maximum cosine* and *prompt cosine* measures exceed the thresholds. For *bad faith* essays, the average rate of false negatives is approximately 26%.[5] A newer prompt-specific method has been developed recently that yields better performance. For proprietary reasons, we are unable to present the methods in this paper. For this proprietary method, the rate of false positives is 5%, and the rate of false negatives is 24%. For the *bad faith* essay data, the false negative rate was 1%. Unfortunately, this new and improved method still requires the topic-specific sets of human-scored essays for training.

## 2.2 Identifying Off-Topic Essays Using CVA & No Topic-Specific Training Data

An alternative model for off-topic essay detection uses content vector analysis (CVA)[6], and also relies on similarity scores computed between new essays and the text of the prompt on

---

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

[3] The formula for calculating the *z-score* for an new novel essay is: $z\text{-}score = (value - mean) \div SD$

[4] See *Data Section 2.2.2* for descriptions of the data sets.

[5] We cannot compute a false positive rate for the *bad faith* essays, since they are not written to any of the 36 topics.

[6] During the course of this study, we have experimented with applying another vector-based similarity measure to this problem, namely Random Indexing (RI) ([18]). Our results indicated that CVA had better performance. We speculate that the tendency of Random Indexing (RI), LSA, and other reduced-dimensionality vector-based approaches to assign higher similarity scores to texts that contain similar (but not

which the essay is supposed to have been written. Unlike the method described in Section 2.1, 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 an on-topic/off-topic 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.

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

### 2.2.1 Content Vector Analysis

The similarity scores needed for this method of off-topic essay detection are calculated by content vector analysis. CVA is a vector-based semantic similarity measure, in which a content vector is constructed for the two texts to be compared, and their similarity is calculated as the cosine of the angle between these content vectors ([19]). 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, 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 $(1+\log(tf))\times\log(D/df)$, 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 our model are taken from an external source, however. Ideally, we could calculate how many documents each term appears in from a large corpus of student essays. Unfortunately, we do not have a sufficiently large corpus available to us, so instead, we use document frequencies derived from the TIPSTER collection ([11]), making the assumption that these document frequency statistics will be relatively stable across genres.

### 2.2.2 Data

Two sets of data are used for this experiment: *unexpected topic* essays and *bad faith* essays. 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 6th through 12th grade. There is an average of 5 topics per grade. These data are all good faith essays that were written to the expected topic.[7] The data used to evaluate the detection of *bad faith essays* were a set of 732 essays for which a human reader has assigned a score of '0'. These 732 essays were extracted from a larger pool of approximately 11,000 essays that had received a score of '0.' Essays can receive a score of '0' for a number of reasons, including: the essay is blank, the student only types his or her

---

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.

[7] 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.

name into the essay, the student has only cut-and-pasted the essay question, or the essay is off-topic. Of the 11,000, we determined that this set of 732 were *bad faith, off-topic* essays, using an automatic procedure that identified an extremely low percentage of words in common between the test question and the essay response. These essays were taken from a different population than the 6th through 12th grade essays. These were from a graduate school population. In addition, none of the essay questions these essays were supposed to respond to were the same as the 36 test questions in the 6th to 12th grade pool of essay questions. We also manually read through this set of 732 essays to ensure that they were *bad faith* essays as opposed to the *unexpected topic* type.

## 2.3 Evaluation & Results

### 2.3.1 Unexpected Topic Essays

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 discrete 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 which were 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. 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 1. 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 essays and to prompt A, while all essays on prompt B are compared to the 34 reference essays and to 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 1. Finally, all essays on prompt A are compared to the 34 reference essays and to prompt B, while all essays on prompt B are compared to the 34 reference essays and to prompt A. This allows us to generate the false negative rates for the training set in Figure 1.

Figure 1 shows the tradeoff 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. 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 represent only a moderate degradation in performance compared to the supervised methods described in *Section 3.1*, but are achieved without the use of labeled training data.
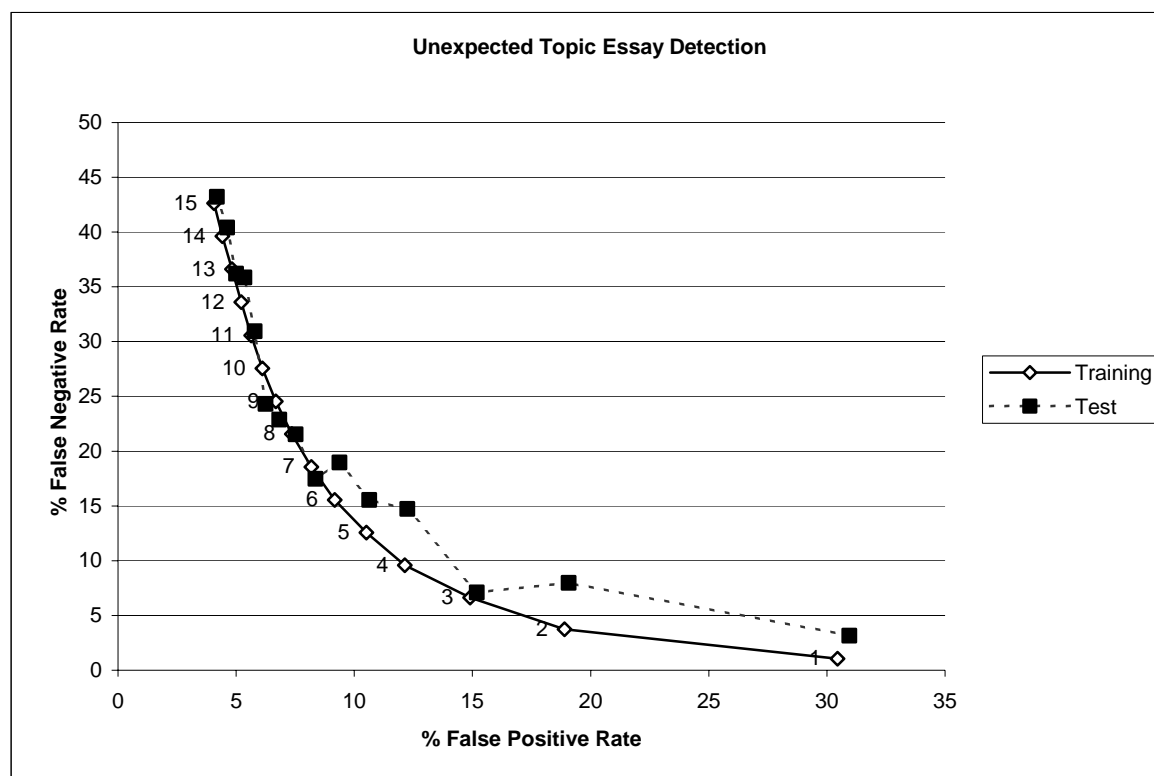


**Figure 1: Performance of CVA-based model in predicting unexpected-topic essays**

### 2.3.2 Bad Faith Essays

For identifying *bad faith* essays, it is more appropriate to use a similarity cutoff because we do not expect these essays to share much vocabulary with any prompt. These are the worst-case off-topic essays, where no attempt was made to answer any kind of essay question.

To evaluate this simple model for detecting bad-faith essays, we generated similarity scores each of the 36 prompts and each of the 732 known bad-faith essays. All essays whose CVA similarity scores with a prompt fell below a cutoff value were correctly identified as bad-faith. If we then count the essays from this set that were not identified as bad-faith, this gives us the false negative rates in Figure 2. Using the same cutoff values, we evaluated how many of the on-topic essays for each of the 36 prompts would be identified as bad-faith by this method. This resulted in the false positive rates in Figure 2. Performance outcomes for the *unexpected topic* and the *bad faith* essay detection evaluations are reported in Figure 2, for a range of similarity cutoff values. **Similarity cutoff values label selected points on the continuous graph that shows the tradeoff between false positives and false negatives. The best cutoff value for our application is probably around .005, which gives us a false positive rate of 3.7% and a false negative rate of 9.2%.**
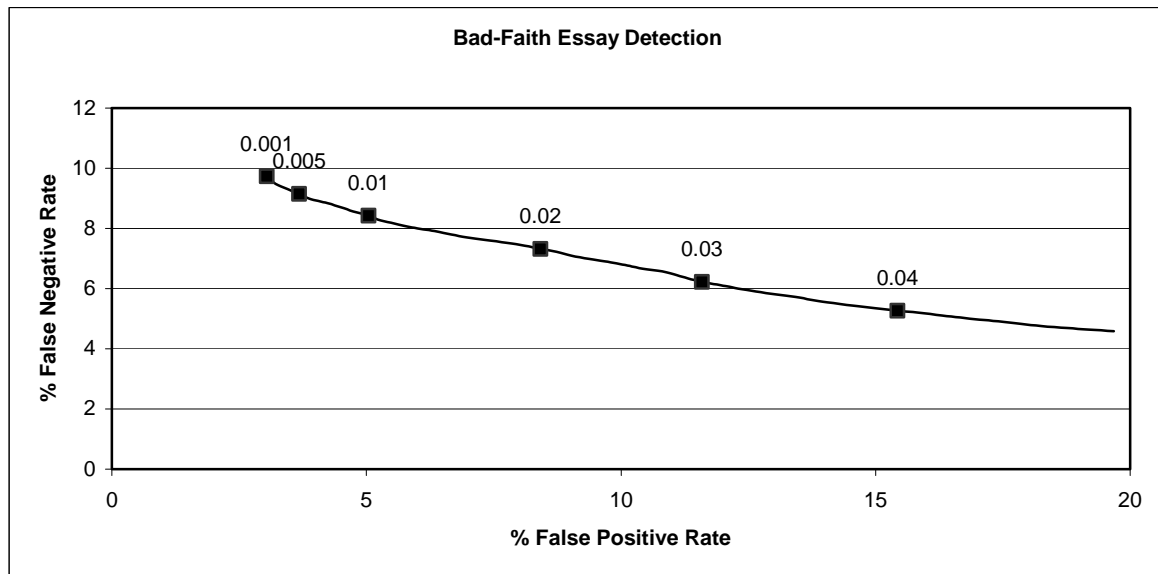
**Bad-Faith Essay Detection**



**Figure 2: Performance of CVA-based model in predicting bad-faith essays**

## Discussion and Conclusions

*Criterion*[SM] is an on-line essay evaluation service that has over 500,000 subscribers. Currently, the system has only a supervised algorithm for detecting off-topic essays input by student writers. Since this supervised method requires 200 – 300 human-scored essays to train each new essay question, the application can not provide feedback about off-topic writing for topics entered on-the-fly by instructors, and 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*. This new method uses content vector analysis to compare a new essay with the text of the essay to which it is supposed to be responding (target prompt), as well as a set of additional essay question texts. Based on these comparisons two procedures are applied. One procedure evaluates if the essay is on topic using the value between a new essay and the target prompt. If this value is amongst the highest CVA values, as compared to the values computed between the same essay and all other prompts, then the essay is on topic. If the essay-prompt comparison shows that the CVA value is not amongst the highest, then this method indicates with similar accuracy to the supervised method, that the essay is off topic, and also an *unexpected topic* essay. In the second procedure, a CVA value is selected that represents a lower threshold, based on a set of CVA essay-prompt comparisons. This lower threshold value represents an essay-prompt comparison in which the two documents contain little word overlap. If the CVA value computed between a new essay and the target prompt is equal to or lower than the pre-set lower threshold, then this is indicative of a *bad-faith* essay. In future work, we plan to look at additional kinds of off-topic writing.

## Acknowledgements

## References

[1] Allan, J. Carbonell, J., Doddington, G., Yamron, J. and Yang,Y. "Topic Detection and Tracking Pilot Study: Final Report." Proceedings of the Broadcast News Transcription and Understanding Workshop, pp. 194-218, 1998.

[2] Attali, Y., & Burstein, J.  (2004, June). Automated essay scoring with e-rater V.2.0. To be presented at the Annual Meeting of the International Association for Educational Assessment, Philadelphia, PA.

[3] Billsus, D. & Pazzani, M. (1999). A Hybrid User Model for News Story Classification, Proceedings of the Seventh International Conference on User Modeling (UM '99), Banff, Canada, June 20-24, 1999.

[4] Burstein, J.  et al. (1998). Automated Scoring Using A Hybrid Feature Identification Technique. Proceedings of 36th Annual Meeting of the Association for Computational Linguistics, 206-210. Montreal, Canada

[5] Burstein, J. et al (2004). Automated essay evaluation: The Criterion online writing service. AI Magazine, 25(3), 27-36.

[6] Burstein, J. (2003) The e-rater® scoring engine: Automated essay scoring with natural language processing. In Anonymous (Eds.), Automated essay scoring: A cross-disciplinary perspective. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

[7] Chapman WW, Christensen LM, Wagner MM, Haug PJ, Ivanov O, Dowling JN, Olszewski RT.  (in press). Classifying Free-text Triage Chief Complaints into Syndromic Categories with Natural Language Processing.  Artificial Intelligence in Medicine.

[8] Cohen, William W., Carvalho Vitor R., & Mitchell, Tom (2004): Learning to Classify Email into "Speech Acts" in EMNLP 2004.

[9] Elliott, S. 2003. Intellimetric: From Here to Validity. In Shermis, M., and Burstein, J. eds. Automated essay scoring: A cross-disciplinary perspective. Hillsdale, NJ: Lawrence Erlbaum Associates.

[10] Foltz, P. W., Kintsch, W., and Landauer, T. K. 1998. Analysis of Text Coherence Using Latent Semantic Analysis. Discourse Processes 25(2-3):285-307.

[11] Harman, Donna. 1992. The DARPA TIPSTER project. SIGIR Forum 26(2), 26-28.

[12] Hripcsak, G., Friedman, C., Alderson, P. O., DuMouchel, W., Johnson, S. B. and Clayton, P. D. (1995). Unlocking Clinical Data from Narrative Reports: A Study of Natural Language Processing Ann Intern Med, 122(9): 681 - 688.

[13] Joachims, T. (2002). Optimizing Search Engines Using Clickthrough Data, Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD).

[14] Larkey, L. 1998. Automatic Essay Grading Using Text Categorization Techniques. Proceedings of the 21st ACM-SIGIR Conference on Research and Development in Information Retrieval, 90-95. Melbourne, Australia.

[15] McCallum, Andrew, Nigam, Kamal, Rennie, Jason and Seymore, Kristie.  Building Domain-Specific Search Engines with Machine Learning Techniques. AAAI-99 Spring Symposium.

[16] Page, E. B. 1966. The Imminence of Grading Essays by Computer. Phi Delta Kappan, 48:238-243.

[17] Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E. 1998. A Bayesian Approach to Filtering Junk E-Mail. In Learning for Text Categorization: Papers from the 1998 Workshop. AAAI Technical Report WS-98-05.

[18] Sahlgren, Magnus. 2001. Vector-based semantic analysis: Representing word meanings based on random labels. In Proceedings of the ESSLLI 2001 Workshop on Semantic Knowledge Acquisition and Categorisation. Helsinki, Finland.

[19] Salton, Gerard. 1989. Information Retrieval: Data Structures and Algorithms. Reading, Massachussetts: Addison-Wesley.

[20] Wilcox AB, Hripcsak G. The role of domain knowledge in automating medical text report classification. J Am Med Inform Assoc 2003;10:330–8.