Detection of Off-topic Responses in Automated Essay Scoring Using Semantic Similarity Measures

Anonymous NAACL submission

Abstract

This paper evaluates the performance of five methods for measuring semantic similarity in the identification of off-topic responses in the context of automated essay scoring, using supervised machine learning models. Prior work has examined the performance of some of these methods. This study compares the performance of these methods across various types of data to understand their performance and how robust they are across these types of data. Specifically, these methods are tested on artificially prepared data with varying proportions of off-topic responses, and operational data at varying levels of inclusion of other types of non-attempt responses. It is seen that Latent Semantic Analysis is largely effective at capturing off-topic responses across these various types of data, and that feeding in semantic vectors as features as opposed to computed similarity scores may work better for operational data.

1 Introduction

The identification of off-topic responses is an important non-attempt classification activity often included in the pre-scoring pipeline of automated essay scoring (AES) systems. While the definition of off-topic tends to vary by application and testing program, in a narrow sense, it refers to any partially or even well-formed essay that is not relevant to the topic elicited by the writing prompt (Higgins et al., 2006). Examinees may submit off-topic responses when they do not understand the question, or when they are incapable or unwilling to formulate a relevant answer.

Off-topic flags are just one type of non-attempt classification performed in AES systems; other types include blank, repetitive, non-English and illegible. Depending on the testing program, some of these other non-attempts are sometimes subsumed under off-topic classification. The variable and sometimes wide scope definitions of off-topic demand that approaches to off-topic detection assume a specific definition, or account for each of these variable definitions in the process of filtering through responses. Building on Higgins et al.'s definition, for the purposes of this study, an off-topic response is a non-blank legible response that is not relevant to the topic elicited by the writing prompt and is not overly repetitive or written in a language other than English.

The focus of this study was to build supervised models to perform off-topic identification using topic modeling methods and semantic similarity, and to compare the performance of the various methods of measuring semantic similarity in different data settings. In operation, different proportions of off-topic responses may occur, depending on the task at hand and the nature of test-takers. Operational data also tends to contain responses that may not be scorable by automated engines in various ways. The motivation behind this study is to determine which techniques are most robust to these variations in data.

Generally, techniques to compute measures of semantic similarity are used to assess topical relevance. For the purposes of this study, the following approaches were used: tf-idf, Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), doc2vec and Word Movers Distance (WMD).

Two research questions were considered for this study:

1. How do the topic modeling and semantic vector methods perform on data sets that have different on-topic to off-topic class ratios?

2. How do the topic modeling and semantic similarity methods perform on datasets with varying levels of inclusion of other non-attempt responses, i.e. a dataset that includes blank, repetitive, non-English and illegible responses vs. one

that does not? And across these varying levels, does it make a difference if topical relevance scores are used as features for the machine learning classifier versus feeding in semantic vectors directly?

Two sets of data were used in this study, one publicly available, originally unlabeled dataset, used to address the first research question, and one operational, labeled dataset, used to address the second.

The structure of the rest of this paper is as follows: section 2 of this paper details a literature review of the various approaches of past researchers on off-topic classification, section 3 describes the datasets used, section 4 gives an overview of the five methods used to ascertain semantic similarity and section 5 describes the experimental design. This is followed by sections detailing the results of the experiments and a concluding summary.

2 Literature Review

The purpose of this section is to present the findings from the relevant literature regarding the differences in performance among of the aforementioned methods and to give an overview of the various approaches that have been used to define and select appropriate methods.

A study by Higgins et al. (2006) compared the performance of supervised and unsupervised models both derived from a pool of responses, from a set number of prompts. It was concluded that the supervised models performed better, and that they are preferable when topic-specific training data is available. The models they employed were to detect several types of off-topic essays: empty responses, banging-on-the-keyboard essays, copied-prompt essays, unexpected-topic essays and irrelevant musings essays. Unexpected-topic and irrelevant musings responses adhere to the narrow definition of off-topic responses provided in the previous section possibly well-formed essays that do not respond to the expected test question with the difference being that irrelevant-musings essays are a result of the student deliberately not choosing to answer the question, whereas unexpected-topic essays may be the result of an examinee copy-pasting a previously prepared essay.

Yoon et al. (2017) used pre-trained publicly available word embeddings to assess document similarity in various ways to identify off-topic responses. They started with a baseline tf-idf vector space model (VSM) constructed for question and response. The other models constructed included averaged and idf-weighted word embeddings vectors, Word Movers Distance (WMD) and a Siamese Convolutional Neural Network (Siamese-CNN) model. The WMD and CNN models worked well, and it was concluded that combination of all these models worked the best. The present study took insights from this research and used document embeddings and WMD and assessed their performance.

Rei and Cummins (2016) did a study investigating unsupervised methods using neural embeddings and word overlap for assessing the topical relevance of individual sentences, with the assumption that sentence-based relevance scores can be used to determine coherence in an essay. They found that weighted embeddings, where the weights were learned to optimize sentence-level vector similarity, performed the best for short prompts, but for longer ones the tf-idf model tended to work well. While sentence-level scores were not generated in the present study, these insights were used in determining the methods, with tf-idf being selected as a baseline method and embeddings being used at the document level.

3 Data

The dataset used to answer the first research question was constructed from the publicly available Automated Student Assessment Prize (ASAP) dataset from The Hewlett Foundation. This data was not labeled for any non-scorable responses. It consisted of eight essay sets with each essay set associated with a different prompt text. The n-counts of the essay sets are given in Table 1; these numbers are from the training data as provided by the foundation, which was used for training, validation and testing in this study. The data was prepared for training by randomly shuffling the responses and assigning them to prompts; if a response was in sample that was assigned to a prompt other than the one it was written for, the response was classified as off-topic.

The dataset used to answer the second research question was a labeled operational dataset from a set of assessments designed for third to eighth grade students. This dataset consisted of 1358 responses to a Constructed Response (CR) writing item. This data contained an off-topic flag, attributed to 372 responses (27.4% of the total data).

Essay set	n-count
1.0	1783.0
2.0	1800.0
3.0	1726.0
4.0	1772.0
5.0	1805.0
6.0	1800.0
7.0	1569.0
8.0	723.0
Total	12978.0

Table 1: N-counts of essay sets in ASAP data.

Flag	n-count
Off-topic (OT)	372.0
Blank (BL)	42.0
Void (VO)	60.0
Illegible (IL)	157.0
Non-English (NE)	39.0
No flag	688.0
Total	1358.0

Table 2: N-counts of flags in operational data.

Additionally, it contained other flags to indicate a non-attempt, namely for blank, illegible, non-English and void responses. The n-counts of the flags in this dataset are given in Table 2.

4 Methods for Measuring Semantic Similarity in Automated Essay Scoring

The approach to model building for this study consisted of building a corpus of documents from the pool of responses and prompts, and then using this corpus to map every document to a vector space. This mapping was done using methods of varying complexity, and are described as follows.

4.1 tf-idf

Term frequency inverse document frequency is a common method to weight terms in information retrieval and document classification. The multiplication by idf attempts to scale up the effect of frequency for terms that only occur in a small proportion of documents in the corpus.

As inputs into this formula a normalized tf-idf matrix is generated from the corpus of documents, and cosine similarity is computed between every response and prompt vector. This similarity metric is then used as a feature in the classifier.

4.2 Latent Semantic Analysis

Latent Semantic Analysis or LSA is a technique in which the dimensionality of a tf-idf matrix is reduced by means of truncated singular value decomposition (SVD) (Islam and Hoque, 2010). This reduced matrix acts as a representation of the latent semantic structure of the corpus, and contains a vector of N dimensions for each document, where N is the number of components defined during SVD. This compact topic vector can then be used to compute cosine similarity.

4.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is used widely as a topic modeling method. In this method, documents are represented as mixtures of topics that are represented by words with certain probabilities. LDA attempts to learn the topic representation for every document in a corpus and the words associated each topic for a fixed number of K topics (Blei et al., 2003).

The similarity between two LDA topic distributions can be computed by calculating the Jenson-Shannon divergence, a common method of measuring similarity between two probability distributions, and as implemented in AES studies employing LDA such as that of Dascalu et al. (2017).

4.4 Doc2vec

Doc2vec is a method presented by Le and Mikolov (2014), centered around representing documents numerically, using an extension of word2vec. Word2vec attempts to learn numeric representations for words from context that are able to encapsulate relations between words such as synonymy or analogy. Doc2vec is a document-level extension of word2vec.

The standard Paragraph Vector Distributed Memory (PV-DM) method of doc2vec, where a paragraph ID or document-unique feature vector is trained, was used in the present study, and cosine similarity was computed between response and prompt as a feature.

4.5 Word Mover's Distance

Word Movers Distance (WMD) is a distance measure between two documents based on word embeddings. For any two word vectors, the Euclidean distance between them can be calculated using a special implementation of Earth Movers Distance (Kusner et al., 2015).

5 Experimental Design

This study was performed on two datasets as previously mentioned the originally unlabeled ASAP and the labeled operational data and the experiment involved feeding features extracted using the abovementioned methods to perform machine learning classification and predict whether the response is off-topic. This section is divided into two parts, one for each dataset, and is intended to be a description of the manner in which the semantic extraction methods were applied on each dataset.

5.1 ASAP data

The originally unlabeled ASAP data was used to perform a preliminary assessment of each method for semantic similarity, and also to answer the first research question regarding the performance of the methods in different settings of on-topic to off-topic label ratio.

This data, after shuffling of responses to prompts and adding the off-topic flag, contained on-topic to off-topic responses in the ratio of 1:8. This original ratio was altered to observe the performance of each of the semantic vector extraction methods in various data settings. With random sampling, datasets with ratios 2:1, 5:1 and 20:1 of on-topic to off-topic responses were constructed. A 20:1 ratio was included as it is a more realistic ratio to be expected in an operational setting.

A Support Vector Machine classifier, with grid search parameter tuning and cross-validation, received the features extracted and performed classification. The grid search chose the best kernel, from either linear or RBF, the value of C, a parameter that accounts for trade-off between misclassification of training examples versus the simplicity of the model, and the value of gamma, which defines how far the influence of a single training example reaches for an RBF kernel.

5.2 Operational data

The operational data was used to address the second research question regarding how the semantic similarity methods perform on data with varying definitions of off-topic, and how the nature of the features affect the performance of the classifier.

The responses in the operational data were found to be highly varied in their scope - containing a significant number of gibberish or non-English responses, repetition and bad-faith responses/refusals, i.e. not just unexpected topic responses - whether they were labeled off-topic or not.

In order to obtain a clearer picture of model performance, the models were applied on this dataset at two levels, to eliminate noise from other non-attempt responses:

- firstly, on the entire dataset
- secondly, with responses containing non-attempt flags other than the off-topic flag removed

6 Results

The results are presented in two sections, the first describing the results of the analysis completed for the first research question using the ASAP dataset, and the second describing the results of the analysis completed for the second research question using the operational dataset.

6.1 Varying ratios of on-topic to off-topic data - ASAP dataset

The results for the four on-topic to off-topic ratio settings on the ASAP data, for the five semantic similarity methods and some combinations thereof, are presented in Tables 3 to 6. The metrics presented are accuracy, precision, recall, F1 score and Matthews Correlation Coefficient (MCC). MCC is a representative measure of the quality of binary classifications that ranges from -1 to +1, and is only high when the classifier performs well on both negative and positive elements (Boughorbel et al., 2017).

Across all data settings, LSA was observed to be the most effective as a lone method, with MCC values ranging from 0.85 to 0.93. In the 5:1 and 20:1 on-topic to off-topic data settings, tf-idf and LDA did not detect any off-topic responses yielding MCCs of 0, but LSA still performed consistently well. Varying combinations of LSA, LDA and doc2vec produced MCC values greater than or equal to LSA MCCs for all off-topic ratio samples, except for the 20:1 ratio. For the 20:1 on-topic to off-topic data setting, doc2vec performed poorly, with an MCC of 0.45, and therefore did not improve the combination methods. These methods yielded MCCs ranging from 0.75 to 0.88, as compared to an MCC value of 0.91 for LSA.

It can therefore be deduced that LSA appears to be the most effective in classifying responses as

Method	Acc.	Prec.	Recall	F1 Score	MCC
tf-idf	0.96	0.96	0.86	0.90	0.82
LSA	0.98	0.98	0.95	0.96	0.92
LDA	0.97	0.95	0.91	0.93	0.86
doc2vec	0.96	0.92	0.89	0.90	0.81
WMD	0.91	0.82	0.72	0.75	0.52
LSA + LDA	0.99	0.99	0.96	0.97	0.95
LSA + doc2vec	0.99	0.98	0.98	0.98	**0.96**
LSA + LDA + doc2vec	0.99	0.99	0.97	0.98	**0.96**

Table 3: Performance on ASAP data with an on-topic to off-topic ratio of 1:8.

Method	Acc.	Prec.	Recall	F1 Score	MCC
tf-idf	0.93	0.90	0.92	0.91	0.82
LSA	0.97	0.97	0.96	0.96	0.93
LDA	0.86	0.84	0.81	0.82	0.64
doc2vec	0.94	0.95	0.89	0.91	0.84
WMD	0.76	0.71	0.70	0.71	0.41
LSA + LDA	0.98	0.97	0.97	0.97	0.94
LSA + doc2vec	0.98	0.98	0.98	0.98	0.96
LSA + LDA + doc2vec	0.99	0.99	0.98	0.99	**0.97**

Table 4: Performance on ASAP data with an on-topic to off-topic ratio of 2:1.

Method	Acc.	Prec.	Recall	F1 Score	MCC
tf-idf	0.88	0.44	0.50	0.47	0.0
LSA	0.97	0.94	0.91	0.93	0.85
LDA	0.88	0.44	0.50	0.47	0.0
doc2vec	0.97	0.96	0.87	0.91	0.83
WMD	0.90	0.95	0.58	0.62	0.39
LSA + LDA	0.97	0.94	0.91	0.93	0.85
LSA + doc2vec	0.99	0.99	0.94	0.96	**0.93**
LSA + LDA + doc2vec	0.99	0.97	0.96	0.96	**0.93**

Table 5: Performance on ASAP data with an on-topic to off-topic ratio of 5:1.

off-topic. Doc2vec is also consistently effective, but only in the presence of adequate data from both the on- and off-topic classifications. Combinations of the LSA and doc2vec methods may yield better results when both methods perform well individually.

Method	Acc.	Prec.	Recall	F1 Score	MCC
tf-idf	0.95	0.47	0.50	0.49	0.0
LSA	0.99	0.99	0.92	0.95	**0.91**
LDA	0.95	0.47	0.50	0.49	0.0
doc2vec	0.96	0.90	0.63	0.69	0.45
WMD	0.96	0.48	0.50	0.49	0.0
LSA + LDA	0.99	0.99	0.90	0.94	0.88
LSA + doc2vec	0.98	0.99	0.82	0.88	0.79
LSA + LDA + doc2vec	0.98	0.99	0.79	0.86	0.75

Table 6: Performance on ASAP data with an on-topic to off-topic ratio of 20:1.

6.2 Varying inclusion of other non-attempt responses - Operational dataset

The results on the operational dataset are presented below for the two filtering levels - original and with non-off-topic flags removed - in Tables 7 and 8. It was observed with this data that there were differences in the results depending on how features were fed into the classifier: whether as similarity scores between vectors, or as the extracted semantic vectors directly. There are therefore two results reported for each method one in which similarity scores were computed from extracted question and response semantic vectors were fed in as features, and other in which the semantic vectors of the responses were fed in as features. Only the best-performing methods as ascertained in the previous section, LSA and doc2vec, were used in this analysis.

These results indicate that directly feeding in semantic vectors as features resulted in models that performed far better than when similarity scores were used as features. While the semantic vectors produced better metrics, the results are not as good as the ASAP dataset metrics. In particular, for the operational dataset, using similarity scores as features produced negligible (defined here as having an MCC of 0 to 0.2) or negative correlation, and was not effective in classifying responses as off-topic. Given that this data had inconsistent off-topic flagging, feeding semantic vectors directly might have helped the model to perform better as there was more data from which to learn and generalize. The best results are seen with feeding in LSA vectors as features for the second definition of off-topic data in Table 8, with an MCC of 0.72,

Method	Similarity Scores as Features					Semantic Vectors as Features				
	Accuracy	**Precision**	**Recall**	**F1 score**	**MCC**	**Accuracy**	**Precision**	**Recall**	**F1 score**	**MCC**
LSA	0.73	0.69	0.55	0.52	0.19	0.81	0.81	0.69	0.72	**0.49**
doc2vec	0.69	0.35	0.48	0.41	-0.11	0.71	0.62	0.59	0.59	0.21
LSA + doc2Vec	0.71	0.61	0.55	0.53	0.14	0.79	0.77	0.67	0.69	0.43

Table 7: Performance on original operational data.

Method	Similarity Scores as Features					Semantic Vectors as Features				
	Accuracy	**Precision**	**Recall**	**F1 score**	**MCC**	**Accuracy**	**Precision**	**Recall**	**F1 score**	**MCC**
LSA	0.60	0.51	0.50	0.49	0.01	0.87	0.90	0.82	0.84	**0.72**
doc2vec	0.60	0.55	0.55	0.55	0.09	0.76	0.74	0.75	0.74	0.49
LSA + doc2vec	0.67	0.61	0.58	0.57	0.18	0.88	0.89	0.84	0.86	**0.72**

Table 8: Performance on operational data with data containing non-off-topic flags removed.

as highlighted.

A qualitative analysis of the predicted flags for the test data revealed that without the removal of other flags, the test responses correctly classified as off-topic tended to be shorter responses with more gibberish and refusals. As more non-off-topic non-attempt categories were filtered out of the input data, more of the correctly classified off-topic responses conformed to the intended specific definition of an off-topic response. This suggests that off-topic detection based on semantic similarity would work best at the end of a well-engineered pre-scoring filtering pipeline, after other non-scorable filters have been applied.

7 Conclusions

In this study, two datasets were used to study the performance of five methods of measuring semantic similarity in building supervised off-topic detection models in the context of automated essay scoring. The first dataset was used to answer the research question about how these methods perform on datasets with varied on-topic to off-topic class ratios. On this dataset, the performance of these methods by computing similarity scores between responses and prompts using them was compared across various class ratios. It was found that while methods employing tf-idf scores and LDA performed poorly at imbalanced class ratios, LSA was the most effective semantic similarity method for off-topic detection in skewed-label settings, with doc2vec also performing well with adequate data.

These insights were extended to the second dataset of which subsets were created after filtering out other types of non-scorable responses, in order to address the second research question concerning performance on datasets with varying levels of inclusion of other non-attempt responses. It was seen that LSA performed well on these subsets, far better than on the original data with a broader definition of off-topic, when LSA output vectors were used directly as training data instead of similarity scores, suggesting that the models may work best on data that has been passed through other non-attempt filters and therefore contains less of the associated noise. It was also shown that using semantic vectors directly as input features rather than using similarity scores may boost performance.

Future work could extend these insights to other operational datasets with different types and scope of labeling for non-scorable responses. Optimizing the pre-scoring flagging pipeline in order to better filter out other types of non-scorable responses may help to ensure that cleaner data is available to the supervised off-topic detection models to allow them to perform better, and could be experimented with further. Using pre-trained word embeddings might help to improve the performance of word and document embedding semantic extraction and similarity methods, and is worth studying.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. 2017. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PloS one*, 12(6):e0177678.

Mihai Dascalu, Wim Westera, Stefan Ruseti, Stefan Trausan-Matu, and Hub Kurvers. 2017. Readerbench learns dutch: Building a comprehensive automated essay scoring system for dutch language. In

International Conference on Artificial Intelligence in Education, pages 52–63. Springer.

Derrick Higgins, Jill Burstein, and Yigal Attali. 2006. Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, 12(2):145–159.

Md Monjurul Islam and ASM Latiful Hoque. 2010. Automated essay scoring using generalized latent semantic analysis. In *2010 13th International Conference on Computer and Information Technology (ICCIT)*, pages 358–363. IEEE.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Marek Rei and Ronan Cummins. 2016. Sentence similarity measures for fine-grained estimation of topical relevance in learner essays. *arXiv preprint arXiv:1606.03144*.

Su-Youn Yoon, Chong Min Lee, Ikkyu Choi, Xinhao Wang, Matthew Mulholland, and Keelan Evanini. 2017. Off-topic spoken response detection with word embeddings. In *INTERSPEECH*, pages 2754–2758.