Identifying Thesis Statements in Student Essays: The Class Imbalance Challenge and Resolution

Fattaneh Jabbari, Mohammad H. Falakmasir, Kevin D. Ashley

Intelligent Systems Program, University of Pittsburgh
{faj5, mhf11, ashley}@pitt.edu

Abstract

A thesis statement or controlling idea is a key component of the Common Core State Standards of writing from grade 6 to grade 12. We developed a machine learning model to identify thesis statements in students' essays in order to focus peer-reviewers on commenting on the presence and quality of an author's thesis statement. Identifying thesis statements in essays can be considered as a classification task in which a classifier is trained to predict whether a sentence is a thesis statement or not based on the features extracted from the sentence. However, the number of sentences in the thesis class is usually much lower than those in the not thesis class. Our initial model could not deal adequately with the challenge of class imbalance; there were too few instances of thesis statements from which to learn. Our subsequent model employs synthetic over-sampling in order to address this challenge and improve performance.

1. Introduction

A thesis statement or controlling idea is a key component of the Common Core State Standards of writing from grade 6 to grade 12. In writing a research report, sixth graders are encouraged to do preliminary research, write research questions, gather and evaluate sources, and create note cards in order to think up a thesis statement that sums up their research. In planning an argumentative essay, twelfth graders, are advised to develop a thesis, identify authoritative sources based on the thesis, organize the evidence from different sources in a persuasive manner, and then write the essay.

Although computer-supported peer-review can help students to understand the difference between effective and ineffective thesis statements, reviewers may fail to identify thesis statements in essays. A machine learning model that can automatically identify thesis statements in essays can facilitate the process and provide a scaffold for a reviewer's reflection.

From the viewpoint of natural language processing, one might define the task of identifying theses as a binary classification task that looks for certain patterns based on a set of features extracted from sentences in a set of essays for a particular assignment. However, there may be too few instances of thesis statements in essays to train a classifier. This problem is called class imbalance or skewed data and may lead to poor minority class recognition. Several solutions have been developed to deal with imbalanced datasets at the data and algorithmic levels. Here we explain challenges we encountered in deploying our previous model, focusing primarily on the challenge of class imbalance, and how we resolved it at the data level.

Section 2 provides background on our peer-reviewing approach and its use of our initial model for detecting thesis statements. In Section 3, we explain the challenge of dealing with imbalanced data in educational applications including thesis detection. In Section 4, we describe the solution to the challenge of class imbalance: synthetic over-sampling of minority instances, here thesis statements. Section 5 reports our methodology for integrating synthetic over-sampling and an experiment to compare our two models with results reported in Section 6. Section 7 provides a discussion of the results and conclusions.

2. Background

The study reported in this paper is part of a larger study to scaffold peer review of writing assignments by focusing on the core elements of writing such as thesis statements (Falakmasir et al., 2014). A thesis statement plays a critical role in an argumentative or analytical essay since it conveys the author's opinion about the essay prompt, constructs the framework of the essay, and plays a role in anticipating critiques and counterarguments (Durst, 1987).

Figure 1. (a) Scaffolding when the model is unable to identify the thesis statement. (b) Scaffolding when the model identifies a sentence as thesis candidate.

In our previous work, we built a model to identify thesis statements so that our system could focus reviewers' attention on commenting on the presence and quality of a thesis statement in student work. We applied different feature selection methods in order to pick the top-n that generalize the decision boundary between thesis statements and other sentences. We also assumed that thesis statements are placed in the first paragraph of each essay and used only those sentences for training purposes. Then, we built a model based on the top-n features and performed a pilot study with 35 students in order to investigate the effect of the model on students' peer-reviewing behavior. The students were writing analytical essays in the form of book reviews.

Our model scaffolds peer-reviews as illustrated in Figure 1. If the model is unable to find a thesis sentence in the essay (Figure 1.a) (i.e., no sentence has a probability above the threshold of 0.8) the system asks the reviewer to identify a thesis in the essay, copy it from the essay to the comment box, and provide feedback about it. If the model identifies a sentence as a thesis candidate from the essays (Figure 1.b), the system will draw the attention of the reviewer to the thesis and ask the reviewer whether s/he agrees with the system. Next, the system asks for the reviewer's comments about the thesis.

We hypothesized that this form of intervention serves to draw the attention of the reviewer to the need for critically considering thesis statements while minimizing the effect of the model's lower performance in lower grades. Whether this form of scaffolding will result in confusion among the students is an empirical question that we have to investigate.

The results of our pilot study, however, showed that the model learned a threshold based on our training data that was too high for a new assignment, and the model was unable to identify thesis statements from the essays in most cases. The model identified only 7 thesis statements out of 35 essays. This is despite the fact that peer rating of the essays on a criterion related to thesis quality had high reliability (ICC(C, 5) = 0.77) although the average rating was 4.4 out of 7.

We decided to investigate the issue in more depth by asking two experts to identify and rate the quality of thesis statements from 1 to 4. The average expert rating for thesis statements was 2.3 out of 4 but in 27 essays, the sentence with the highest probability based on our model was the same sentence identified by the expert as the thesis statement. As an example, this is a sentence that was rated 2 out of 4 by the expert: *"Chappie has relationships that negatively influenced him, but he also made strong positive relationships with friends he will never forget."* Our model also identified this sentence as a thesis candidate but since the probability was below the threshold (0.8), the system did not return it as a candidate.

Consequently, we decided to improve the thesis identification model to better distinguish between thesis and non-thesis sentences by balancing the training data distribution rather than performing feature engineering and changing the threshold.

3. The Challenge: Class Imbalance

A major challenge that we encountered in building the new model was class imbalance or skewed data. Dealing with imbalanced data is one of the most important challenges in knowledge discovery and data mining since it arises in many real-world practical applications such as fraud/intrusion detection (Fawcett & Provost, 1996), risk management, text categorization (Dumais et al., 1998), and detecting integration in student essays (Hasting et al., 2012).

A dataset is called imbalanced when one class has many more instances compared to other classes. Many machine learning algorithms assume balanced class distribution in data. Thus they tend to be biased toward the majority class due to over-prevalence. This bias leads to poor performance on predicting samples of the minority class, even though in such applications one is often more interested in class prediction of the minority samples.

Class imbalance is also common in educational contexts. A targeted pedagogical concept like an effective thesis statement or anticipating a counter-argument may be quite rare in student writing even in a larger corpus of es-

says. Class imbalance presents an obstacle to the thesis identification task since, in any given essay only a very small number of sentences, usually one or two, are labeled as a thesis.

One of the earliest attempts to resolve this issue was the research of Burstein et al. (2003a) in the context of automated identification of discourse structures in student essays. They built a probabilistic-based discourse analyzer and used a noisy channel framework in order to assign a set of pedagogically meaningful labels such as: title, introductory material, thesis, etc. to different sentences of the essays. Using the noisy-channel model allowed them to benefit from the ordering of the labels and rhetorical structure of the essays in order to minimize the effect of class imbalance.

In our previous work (Falakmasir et al., 2014), we used the Gini coefficient and an iterative feature selection procedure in order to resolve the class imbalance issue. Starting from 42 features, we ended up with 13 features and built an SVM model that had comparable results with the model of Burstein et al. (2003b).

A more recent example dealing with imbalanced data in the context of intelligent tutoring is the work by Hasting et al. (2014). They have developed a model in order to identify conceptual elements from student essays and infer the causal structure of the essays. They used both over-sampling and synthetic over-sampling (discussed below) in order to deal with the class imbalance problem.

A variety of techniques have been introduced to deal with class imbalance, as summarized in (Chawla et al., 2004) and (He and Garcia, 2009). Application of different sampling strategies is one of the most common ways to deal with class imbalance. In under-sampling, the majority class samples are under-sampled by randomly removing them until the minority class becomes a specified percentage of the majority class. While this helps to improve the sensitivity of the classifier to the minority class, its main drawback is that it can potentially discard useful examples (Liu et al., 2009). Another sampling approach is over-sampling with replication, which simply duplicates minority class samples. This approach may lead to over-fitting and introduce additional computational cost if a data set is large.

In our work, we apply a third powerful tool, the synthetic over-sampling technique (SMOTE), to create new synthetic samples from the minority class (Chawla et al., 2002), in addition to the two other methods. The description of this approach is given in the following section.

4. The Solution: Synthetic Over-sampling

In order to make our previous model more robust, we decided to replace the feature selection step with over-sampling using synthetic data. We used SMOTE to generate the synthetic examples in order to alleviate class imbalance and obtain a better distinction between the sentences that are thesis statements and other sentences within the first paragraph.

As noted, SMOTE is an over-sampling method in which we can create new "synthetic" instances of the minority class by interpolating between existing minority instances rather than simply duplicating the original ones (Chawla et al., 2002). This approach was primarily motivated by (Ha and Bunke, 1997) in which they applied some predefined perturbations such as rotation on the input images of handwritten numerals in order to create more training data.

By contrast, SMOTE operates in the feature space rather than the original data space, i.e. it exploits feature vectors and creates new data points in that space. First, it takes the subset of samples that belong to the minority class. Then for each data point x_i in the minority subset, its K-nearest neighbors are identified. Now a synthetic sample can be generated in the following way:

1. Select one of the K-nearest neighbors randomly (x_k),
2. Take the difference between the corresponding feature vector of x_i and x_k,
3. Multiply this difference by a random number between 0 and 1,
4. Add it to x_i.

The following formula summarizes the procedure described above:

$$x_{new} = x_i + (x_i - x_k) * \mathrm{rand}(0,1)$$

where x_{new} is a new synthesized data point in feature space and $\mathrm{rand}(0,1)$ is a function that generates a random number between 0 and 1. Depending upon the desired amount of over-sampling, the above procedure should be repeated. For example, if 300% more minority samples are needed, for each minority sample x_i, we need to create three artificial data points along the line segments joining x_i to three of its nearest neighbors. This approach thus forces the decision region of the minority class to be larger and more general. Figure 2 shows some synthesized examples in 2-dimensional feature space as gray plus signs. A detailed SMOTE algorithm and pseudo-code are given in (Chawla et al., 2002).

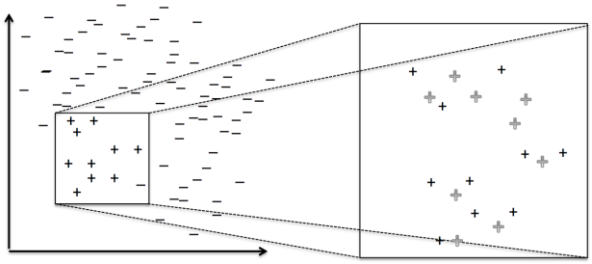

Figure 2. SMOTE synthetic examples

5. Methodology

In this study, we compared three different approaches to tackling the data imbalance problem in our training set in order to improve the thesis identification in student essays. Specifically, we considered random under-sampling and over-sampling with replacement as our baselines. Then we compared their results with the SMOTE synthetic over-sampling. We experimented with two different classifiers for thesis classification: decision tree and support vector machine (SVM) with linear kernel. A decision tree (Safavian and Landgrebe, 1990) is a classifier represented by a tree structure in which leaf nodes denote the class labels and non-leaf nodes help to traverse the tree. SVM (Boser et al., 1992) is a well-known classification technique, which finds a decision boundary that is maximally away from any data points.

5.1. Data Description

Our dataset consists of 432 essays of two high school courses on cultural literacy and world literature, which includes eight distinct writing assignments (Falakmasir et al., 2014). In order to create two distinct training and test sets, 6 assignment prompts with 326 essays were used for training data and the other two assignment prompts with 106 essays were used as the test set.

We created an instruction manual based on the scoring guidelines and sample responses of AP English Language and Composition courses in order to guide six human judges in annotating the dataset. Each essay was coded by at least two human judges. The annotation task was defined to identify candidate thesis statements and rate them on the scale of 1 to 3. Based on criteria in the prepared instruction manual, rating 1 means a vague or incomplete thesis, rating 2 demonstrates a simple but acceptable thesis, and rating 3 indicates a sophisticated thesis statement. Finally, we only used simple (rated 2) and sophisticated (rated 3) thesis candidates as positive instances of thesis statements.

In order to measure inter-annotator agreement, we applied Cohen's Kappa (Fleiss et al., 1969) on both the sentence and essay level. In order to achieve an acceptable agreement, the annotators were asked to re-annotate the data if Kappa was below 0.6. Table 1 shows the details of our training and test sets. As one can see in Table 1, around one percent of sentences in the first paragraphs are labeled as thesis statements in each of the training and test sets. About half of the essays in both sets lack a thesis statement. This shows a severe imbalance in the data.

5.2. Models

For machine learning, we used some basic positional, syntactic, and key term features similar to Burstein et al. (2003b) and Falakmasir et al. (2014). Our positional features included only sentence number in the paragraph. Compared to Falakmasir et al., (2014), we did not use paragraph number and type of paragraph since we are only considering sentences in the first paragraph.

Table 1. Training and test set description

	Training Set	Test Set
Total number of sentences	15359	2722
Number of essays	326	106
Number of essays with thesis	214	70
Number of essays without thesis	112	36
Number of sentences within the first paragraph	3269	409
Number of thesis sentences	218	73

Some of our syntactic features that are mostly defined at the sentence level include prepositional and gerund phrases, and the number of adjectives and adverbs. Key term features include a set of frequent words such as "although", "because", "due to", "led to", and "caused", keywords among the most frequent words of the essay, number of words overlapping with the assignment prompt, and a score-based on Rhetorical Structure Theory (RST) adapted from (Marcu, 1999). After data collection and feature extraction, in order to address the class imbalance we applied three methods to modify the class distribution by using:

- **Method 1:** under-sampling.
- **Method 2:** over-sampling.
- **Method 3:** SMOTE over-sampling.

The third step is to evaluate and compare the results of these methods by experimenting with two different classifiers on both the original and the re-balanced datasets. We used an evaluation protocol visualized in Figure 3. First, the dataset was separated into training and test sets. Next, the enrichment data was generated on the training set. Then the classifiers were built, and finally, the evaluations were applied to the test set.

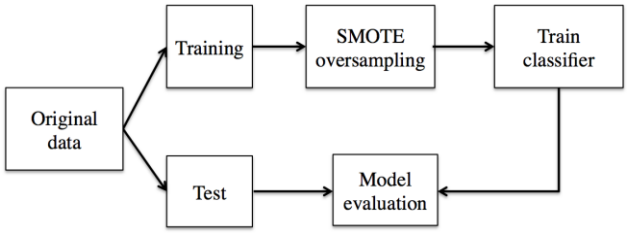

Figure 3. Training and evaluation process

6. Experimental Results

We evaluated our models on both the sentence and essay levels. Tables 2 and 3 illustrate the performance of the classifiers with three different sampling methods. In these tables, P, R, and F indicate precision, recall, and F-measure respectively.

Table 2 demonstrates that SMOTE over-sampling has improved the precision of the SVM classifier in identifying thesis statements from 55% (under-sampling) and 61% (over-sampling) to 64%. Recall is enhanced up to 84% compared to 53% in case of under-sampling and 63% in-case of over-sampling with replacement. Similar improvements are observed in terms of F-measure and Kappa for both classifiers (Table 2).

Table 2. SVM performance on unseen test set – sentence level

		Not Thesis	Thesis	Average	Kappa
Under-sampling	**P**	0.9	0.55	0.84	0.44
	R	0.91	0.53	0.85	
	F	0.91	0.54	0.84	
Over-sampling with replacement	**P**	0.92	0.61	0.87	0.53
	R	0.92	0.63	0.87	
	F	0.92	0.62	0.87	
SMOTE over-sampling	**P**	0.97	0.64	0.91	0.66
	R	0.9	0.84	0.89	
	F	0.93	0.73	0.9	

Table 3 also compares the results of different sampling methods using the decision tree classifier. It shows that while the decision tree classifier with SMOTE over-sampling outperforms the other two sampling methods in terms of all evaluation metrics, it does not perform as robustly and accurately as the similarly equipped SVM classifier. So, for the rest of our experiments, we use the SVM classifier.

Next in our set of experiments, we compared the linear SVM model trained with synthetic data, with our previous best model, which is a fine-tuned linear SVM model used in our previous pilot study. In this experiment, we are only comparing the two models at the essay level because the final deployment of the model on our peer-review system takes a student essay as input and returns a sentence as a thesis candidate. In order to evaluate the models at the essay level, we aggregated the results of the sentence level model in order to predict whether an essay contains a thesis statement or not. Our test set includes 106 essays in which only 70 contain a thesis statement. Table 4 shows the performance of two linear SVM classifiers.

Table 3. Decision tree performance on unseen test set – sentence level

		Not Thesis	Thesis	Average	Kappa
Under-sampling	**P**	0.86	0.37	0.78	0.22
	R	0.88	0.33	0.79	
	F	0.87	0.35	0.78	
Over-sampling with replacement	**P**	0.85	0.38	0.77	0.18
	R	0.92	0.23	0.8	
	F	0.89	0.29	0.78	
SMOTE over-sampling	**P**	0.91	0.51	0.84	0.43
	R	0.89	0.56	0.83	
	F	0.9	0.53	0.84	

Table 4. Comparison of performance of our previous model with new model – essay level

		Not Thesis	Thesis	Average	Kappa
Fine-tuned SVM model	**P**	0.51	*0.78*	0.64	0.3
	R	0.63	*0.68*	0.66	
	F	0.56	*0.73*	0.65	
SVM with SMOTE over-sampling model	**P**	0.68	0.82	0.77	0.48
	R	0.64	0.84	0.77	
	F	0.66	0.83	0.77	

As Table 4 shows, our new SVM model with synthetic over-sampling performs significantly better than our previous fine-tuned SVM model with 82% precision and 84% recall. The Kappa was also improved by 18% using the new model. In this experiment, the SVM model with SMOTE has found thesis statements in 59 out of 70 essays while our previous model was able to find only 48 of them.

In the next experiment, we deployed this new model on the dataset from our pilot study with 35 students in which the students were writing an analytical essay in the form of a book review. In the pilot study, only 16 out of 35 essays have thesis statements, according to our expert-annotated gold standard. Application of SMOTE and the linear SVM resulted in identifying more thesis statements and outperforming the previous model. Results are shown in Table 5.

Table 5. Comparison of two models on pilot study- essay level

		Not Thesis	Thesis	Average	Kappa
Baseline model	**P**	0.54	*0.29*	0.43	0
	R	0.75	*0.13*	0.49	
	F	0.63	*0.18*	0.44	
SMOTE over-sampling model	**P**	0.67	0.59	0.63	0.25
	R	0.63	0.62	0.63	
	F	0.65	0.61	0.63	

7. Discussion and Conclusions

In this work, we extended our previous model to identify thesis statements from students' essays. One of the challenges we encountered was learning an appropriate threshold since the one learned based on our training data was too high for our pilot study; therefore, our original model was unable to find thesis statements of lower quality. In order to achieve a more scalable and robust thesis identification model, we used SMOTE over-sampling to generate the synthetic examples to handle the class imbalance obstacle. We compared SMOTE with random over- and under-sampling approaches using two different classifiers and our experiments showed that SMOTE improves the model performance on the minority class, which is the thesis statements.

Our next step in ongoing research is to deploy our new model and study its effect on the quality of students' writing by using different forms of intervention. For example, we have designed a study to investigate the effects on students' learning and peer-reviewing behavior of always returning a sentence as a candidate thesis statement and of returning only sentences with high probabilities.

References

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). *A training algorithm for optimal margin classifiers.* In Proceedings of the fifth annual workshop on Computational learning theory (pp. 144-152). ACM.

Burstein, J., Marcu, D., & Knight, K. (2003a). *Finding the WRITE stuff: Automatic identification of discourse structure in student essays.* Intelligent Systems, IEEE, 18(1), 32-39.

Burstein, J., & Marcu, D. (2003b). *A machine learning approach for identification thesis and conclusion statements in student essays*. Computers and the Humanities, 37(4), 455-467.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: synthetic minority over-sampling technique*. Journal of Artificial Intelligence Research 16(1), 321–357.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). *Editorial: special issue on learning from imbalanced data sets*. ACM Sigkdd Explorations Newsletter, 6(1), 1-6.

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998, November). *Inductive learning algorithms and representations for text categorization.* In Proceedings of the seventh international conference on Information and knowledge management,148-155, ACM.

Durst, R. (1987). *Cognitive and Linguistic Demands of Analytic Writing*. Research in the Teaching of English, 21(4), 347-376.

Falakmasir, M. H., Ashley, K. D., Schunn, C. D., & Litman, D. J.(2014). *Identifying Thesis and Conclusion Statements in Student Essays to Scaffold Peer Review*. In Intelligent Tutoring Systems (pp. 254-259). Springer International Publishing.

Fawcett, T., & Provost, F. (1997). *Adaptive fraud detection*. Data mining and knowledge discovery, 1(3), 291-316.

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). *Large sample standard errors of kappa and weighted kappa*. Psychological Bulletin, 72(5), 323.

Ha, T. M., & Bunke, H. (1997). *Off-line, handwritten numeral recognition by perturbation method*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 19(5), 535-539.

Hastings, P., Hughes, S., Britt, A., Blaum, D., & Wallace, P. (2014, January). *Toward Automatic Inference of Causal Structure in Student Essays.* In Intelligent Tutoring Systems (pp. 266-271). Springer International Publishing.

He, H., & Garcia, E. A. (2009). *Learning from imbalanced data*. Knowledge and Data Engineering, IEEE Transactions on, 21(9), 1263-1284.

Hughes, S., Hastings, P., Magliano, J., Goldman, S., & Lawless, K. (2012, January). *Automated approaches for detecting integration in student essays*. In Intelligent Tutoring Systems (pp. 274-279). Springer Berlin Heidelberg.

Liu, X. Y., Wu, J., & Zhou, Z. H. (2009). *Exploratory undersampling for class-imbalance learning*. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 39(2), 539-550.

Marcu, D. (1999). *Discourse trees are good indicators of importance in text*. Advances in Automatic Text Summarization, 123-136.

Safavian, S. R., & Landgrebe, D. (1990). *A survey of decision tree classifier methodology.*

(If you use EndNote and it complains "object has been deleted" when you try to generate references, temporarily delete the heading with the copyright notice attached as a footnote, and reinsert it after EndNote finishes.)