Claim Detection and Relationship with Writing Quality

Qian Wan[1], Scott Crossley[1], Laura Allen[2], Danielle McNamara[3]

qwan1@gsu.edu, scrossley@gsu.edu, Laura.Allen@unh.edu, dsmcnama@asu.edu

[1] Georgia State University [2] University of New Hampshire [3] Arizona State University

Wan, Q., Crossley, S. A., Allen, L. K., & McNamara, D. S. (2020). Claim detection and relationship with writing quality. In V. Cavalli-Sforza, C. Romero, A. Rafferty, & J. R. Whitehill (Eds.), *Proceedings of the 13th International Conference on Educational Data Mining (EDM)* (pp. 691-695). Virtual Conference: International Educational Data Mining Society.

Author's Note

This research was supported in part by the Institute for Educational Sciences (IES R305A180261) and the Office of Naval Research (N00014-17-1-2300). Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of IES or ONR.

Claim Detection and Relationship with Writing Quality

Qian Wan
Georgia State University
qwan1@gsu.edu

Scott Crossley
Georgia State University
scrossley@gsu.edu

Laura Allen
University of
New Hampshire
Laura.Allen@unh.edu

Danielle McNamara
Arizona State University
dsmcnama@asu.edu

ABSTRACT

In this paper, we extracted content-based and structure-based features of text to predict human annotations for claims and non-claims in argumentative essays. We compared Logistic Regression, Bernoulli Naive Bayes, Gaussian Naive Bayes, Linear Support Vector Classification, Random Forest, and Neural Networks to train classification models. Random Forest and Neural Network classifiers yielded the most balanced identifications of claims and non-claims based on the evaluation of accuracy, precision, and recall. The Random Forest model was then used to calculate the number, percentage, and positionality of claims and non-claims in a validation corpus that included human ratings of writing quality. Correlational and regression analyses indicated that the number of claims and the average position of non-claims in text were significant indicators of essay quality in the expected direction.

Keywords

argument mining, claim detection, essay quality, natural language processing, automated essay evaluation

1. INTRODUCTION

Argumentative essays include many different discourse units including a thesis statement, main ideas (claims), supporting ideas, and a conclusion (Burstein et al., 2003). Since argumentative essays are important elements in the teaching and assessment of writing, various techniques have been used to identify discourse units including those based on natural language processing (NLP). NLP has been used to automatically identify discourse elements based on the linguistic features that comprise discourse. Previous studies have found that content (i.e., lexical, syntactic, and discourse indicators) and structural features (i.e., the positionality of tokens, sentences, and paragraphs) are effective in the identification of discourse elements (Burstein et al., 1998, 2001a, 2001b, 2003; Lawrence and Reed, 2015; Nguyen and Litman, 2015, 2016; Persing and Ng, 2015; Stab and Gurevych, 2014, 2017). However, most studies have extracted content features at the word-level (unigram) or bigram level (e.g., Stab and Gurevych, 2017), or used indicators that generally occur only as transitional markers either at the beginning or the end of sentences (e.g., Burstein et al., 1998). Less is known about how multi-word n-grams (bigrams and trigrams) and their associated part-of-speech (POS) tags can influence the accuracy of discourse unit identification. Meanwhile, few if any studies, have examined how normalized positions of sentences in paragraphs and in text can predict claims. Lastly, while some studies (e.g., Klebanov et al., 2016) have examined relations between essay quality and the use of discourse structures, these studies have examined relatively small corpora (e.g., test sets of 40 essays) and have not focused on claims, an important discourse element.

Qian Wan, Scott Crossley, Laura Allen and Danielle McNamara "Claim Detection and Relationship with Writing Quality" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 691 - 695

2. PURPOSE STATEMENT AND RESEARCH QUESTION

In this study, we develop NLP approaches to automatically identify claims in structurally-annotated essays using n-grams and POS tags along with positionality data. We compared the identification accuracy of the derived NLP features using different machine learning models and examined the relations between the number (and percentage) of claims and non-claims, their positionality, and human ratings of argumentative essay quality. Two structure-annotated corpora from Stab and Gurevych (2014, 2017) were used as our training (*N* = 329) and testing (*N* = 90) sets. The model with the best performance was used to identify claims and non-claims in a corpus comprising 2269 argumentative essays that had been rated on writing quality. Finally, we conducted correlation and regression analyses to explore the relations between the variables. The research questions that guide this study are as follow:

1. To what extent do (1) the frequency of n-grams (bigrams and trigrams), (2) the frequency of part-of-speech (POS) n-grams (bigrams and trigrams), and (3) positional (structural) information of sentences predict whether the sentence is a claim or not?

2. What are the relations between the number, percentage, and positionality of predicted claims/non-claims in an essay and the quality of the essay?

3. METHOD

3.1 Data

Three corpora were used in the current study. A training and testing corpora were used to train and test the claim detection algorithm, respectively. The claim detection algorithm was then applied to a validation corpus of student essays to calculate the number, percentage, and positionality of claims and non-claims in each essay. The relations of these features to claims (and non-claims) and essay quality was then examined.

3.1.1 Training set

The training corpus was developed by Stab and Gurevych (2017) and was annotated with argument components ("major claim," "claim," and "premises") and the relationships between "premises" and "major claim" or "claims" ("attack" or "support"). The corpus contains 402 argumentative essays written by students on 341 different prompts (e.g. "Will computers replace human power in jobs" and "Should students be taught to compete or cooperate").

The essays were collected from an online writing forum where native and non-native speakers of English could post their argumentative essays and give feedback to each other to help improve writing quality. After removing 73 essays that were duplicated in the testing set, there were 329 essays in the training corpus.

Major claims referred to sentences that directly expressed the general stance of the author that was supported by additional arguments. Claims were the central component of an argument, and premises were reasons that were provided by the author for supporting or attacking a certain claim. Three non-native annotators participated in the annotation process. According to the original study, the overall inter-rater agreement among the three annotators was .72.

3.1.2 Testing set

The testing corpus contained 90 argumentative essays compiled by Stab and Gurevych (2014). The essays were originally collected from the same source as the training set and were annotated by three annotators using the same annotation guidelines as the training set. It is unknown if the same annotators were used. The reported inter-rater reliability was .68.

3.1.3 Validation set

We selected 2269 argumentative essays written by native speakers of English as our validation corpus. The essays were collected in the development of the Writing-Pal (McNamara et al., 2012) from individual participants who composed essays in response to 13 specific prompts. Most of the participants were students ranging in grade levels from 7th to 10th or first-year college students. The participants were asked to respond to a specific prompt, state the degree to which they agreed or disagreed with the statement, and provide supporting evidence and arguments to persuade the readers. The essays in the validation corpus were evaluated by human raters following the scoring rubric used in the SAT (a standardized test used for college admittance in the United States). The SAT rubric evaluated writing in terms of ideas, organization, style, and voice. Raters were asked to assign each essay a quality score between 0-6. Interrater-reliability was greater than Cohen's Kappa .60 and *r* .70. Averages were taken between the two raters. If two raters disagreed by greater than one point on the 6-point scale, they were asked to adjudicate the essay. The average score for the essays was 3.38 and the standard deviation was .91.

3.2 Algorithm Development

Data preprocessing, feature development, application of machine learning models, and the selection of those models were the four major steps in the development of the classification algorithms for the claims and non-claims. We report the first two major steps in the following section and report the application and selection of machine learning models in the results section.

3.2.1 Merge and build standardized structure-annotated sub-corpora

The training and testing corpus were annotated using a framework of three argumentative units ("major claim," "claim," and "premise"). However, in this study we are only interested in distinguishing claims from non-claims. Based on our focus, we merged the tags of "major claim" and "claim" and treated both of them as a larger category of claim. We treated any sentences in an essay that did not fall into the category of claim as a non-claim. We then unified the formats of the two structural annotated corpora by tokenizing the essays into sentence and adding structural tags (claim or non-claim) for each sentence based on the annotation of the original corpora. Further, we extracted all claim sentences from the training corpus to build the claim sub-corpus and extracted all the non-claim sentences to build the non-claim sub-corpus.

3.2.2 N-gram and n-gram POS tokenization

In this study, all of the n-gram and POS n-gram features for model development were extracted only from the training corpus. After the claim and non-claim sub-corpora were built, a Python script was written to tokenize the sentences within each corpus into bigrams and trigrams. Thus, all of the n-grams were extracted on sentence instead of clause levels. Prior to n-gram tokenization, all punctuations within the sentences were removed. Then, all of the characters were set to lowercase and all extra blanks in the sentences were removed from the texts. Stop words (e.g., *of, a, and, the*) were not deleted from the text. The texts were not lemmatized or stemmed.

We used the NLTK (Natural Language Toolkit; Bird et al., 2009) to tokenize the claim and non-claim sub-corpora into bigram and trigram. After the n-gram tokenization, we used the NLTK part-of-speech tagger to label the word class of each word within each sentence in the claim and non-claim sub-corpora. The NLTK pos-tagger labels part-of-speech for each word based on Penn Treebank tagset (Marcus et al., 1993). Prior studies have shown that the overall accuracy of NLTK pos-tagger was 91.33% for Brown Corpus, 89.56% for Treebank Corpus, and 86.45% for NPS Chat Corpus (Yumusak et al., 2014). Once the POS-tagging was completed, we used the NLTK tokenizer to segment the POS-tagged corpora into part-of-speech bigrams and trigrams. For example, the following phrases *should be, would be, can be,* and *will be* would be converted to the same POS n-gram combination: MD (modal) + VB (verb base).

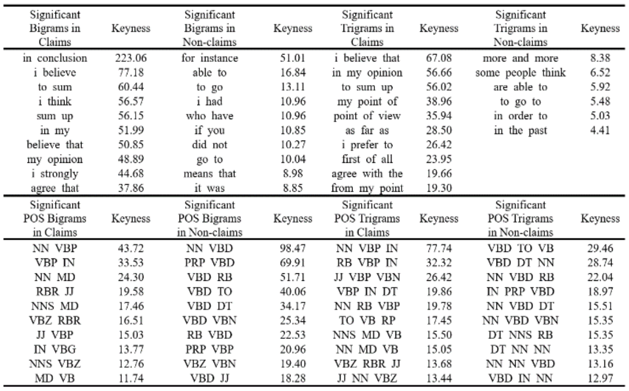
3.2.3 Normalized frequency and Keyness values

We calculated raw frequency and normalized frequency for each bigram, trigram, as well as POS bigram and trigram term in the training corpus (both claim and non-claim sub-corpora). In addition to raw and normalized frequency, keyness value of each n-gram and POS n-gram was also calculated based on the raw frequency data. Keyness value, based on log-likelihood values, provided evidence of whether n-grams and POS n-grams were more common in one corpus compared with the other corpus (Kilgarriff, 2001).

The thresholds for log-likelihood was 3.84 (equivalent to $p < .05$). Specifically, for any n-gram or POS n-gram that appeared in both corpora, if the n-gram or POS n-gram had a log-likelihood value greater than 3.84, we considered it more likely to occur in one corpus over the other. In this study, we wrote a Python script to automatically calculate the Keyness values (log-likelihood values) for all n-grams or POS n-grams that could be found in both claim and non-claim sub-corpora based on Rayson and Garside (2000). In Table 1, we list the top n-grams and POS n-grams with highest keyness values found in claims and non-claims.

In total, we calculated the following indices in the training, testing, and validation corpus, respectively: (1) the frequency of significant n-grams (bigrams and trigrams) in the claims extracted from the training corpus in each sentence; (2) the frequency of significant n-grams in the non-claims extracted from the training corpus in each sentence; (3) the frequency of significant POS n-grams in the claims in each sentence; and (4) the frequency of significant POS n-grams in the non-claims in each sentence. In this way, for each sentence in each corpus, we derived eight indices.

Table 1 Top n-grams with highest keyness values in claims and non-claims



Significant Bigrams in Claims	Keyness	Significant Bigrams in Non-claims	Keyness	Significant Trigrams in Claims	Keyness	Significant Trigrams in Non-claims	Keyness
in conclusion	223.06	for instance	51.01	i believe that	67.08	more and more	8.38
i believe	77.18	able to	16.84	in my opinion	56.66	some people think	6.52
to sum	60.44	to go	13.11	to sum up	56.02	are able to	5.92
i think	56.57	i had	10.96	my point of	38.96	to go to	5.48
sum up	56.15	who have	10.96	point of view	35.94	in order to	5.03
in my	51.99	if you	10.85	as far as	28.50	in the past	4.41
believe that	50.85	did not	10.27	i prefer to	26.42		
my opinion	48.89	go to	10.04	first of all	23.95		
i strongly	44.68	means that	8.98	agree with the	19.66		
agree that	37.86	it was	8.85	from my point	19.30		
Significant POS Bigrams in Claims	Keyness	Significant POS Bigrams in Non-claims	Keyness	Significant POS Trigrams in Claims	Keyness	Significant POS Trigrams in Non-claims	Keyness
NN VBP	43.72	NN VBD	98.47	NN VBP IN	77.74	VBD TO VB	29.46
VBP IN	33.53	PRP VBD	69.91	RB VBP IN	32.32	VBD DT NN	28.74
NN MD	24.30	VBD RB	51.71	JJ VBP VBN	26.42	NN VBD RB	22.04
RBR JJ	19.58	VBD TO	40.06	VBP IN DT	19.86	IN PRP VBD	18.97
NNS MD	17.46	VBD DT	34.17	NN RB VBP	19.78	NN VBD DT	15.51
VBZ RBR	16.51	VBD VBN	25.34	TO VB RP	17.45	NN VBD VBN	15.35
JJ VBP	15.03	RB VBD	22.53	NNS MD VB	15.50	DT NNS RB	15.35
IN VBG	13.77	PRP VBP	20.96	NN MD VB	15.05	DT NN NN	13.35
NNS VBZ	12.76	VBZ VBN	19.40	VBZ RBR JJ	13.68	NN NN VBD	13.16
MD VB	11.74	VBD JJ	18.28	JJ NN VBZ	13.44	VBD IN NN	12.97

3.2.4 *Positional data for sentences*

Beyond n-gram patterns, studies have shown that in argumentative or academic writing, sentence position is an indicator of the structural function of the sentence (e.g., Burstein et al., 1998, 2001a; Biber et al., 2004). In this study, the following raw and normalized positional variables for each sentence in an essay were calculated as potential positional features: (1) the position of the sentence in the whole essay (e.g., if a sentence is the 5th sentence in the essay, the value of this variable would be 5); (2) normalized sentence position in the essay (i.e., equal to the value in [1] divided by the total number of sentences in the essay); (3) the position of the paragraph in which the sentence was located (e.g., if the sentence occurred in the 2nd paragraph of the essay, this value would be 2); (4) normalized paragraph position in the essay (i.e., equal to the value in [2] divided by the total number of paragraphs in the essay); (5) the position of the sentence in the paragraph where the sentence occurred (e.g., if the sentence was the 4th sentence in its paragraph, the value would be 4); and (6) the normalized position of a sentence in a paragraph (i.e., equal to the value in [5] divided by the total number of sentences in the paragraph).

3.3 Validation Study

Our second objective was to examine the relationship of the number/percentage of claims and positional data with the quality (human score) of the essay. To do so, the algorithm (from the final model) to predict the discourse type (claim or non-claim) was applied to each sentence of each essay in the validation corpus. We then calculated the percentages and average position of claim and non-claim sentences in each essay of the validation corpus and used these features to model essay quality to examine the following: (1) correlations between essay quality (represented by human holistic scores of the essays) and the number/percentage and positionality of claims/non-claims in the essay; and (2) the extent to which the number and percentage of claims/non-claims in an essay and sentence positionality predict its quality. In the regression analysis, the number of claims, the number of non-claims, the percentage of claims, the percentage of non-claims in an essay, and sentence positionality were included as the independent variables, while the human score of the essay served as the dependent variable. Prior to analyses, the human scores were checked for normality; multicollinearity ($r < .70$) across all independent variables was checked to ensure the variables developed were unique.

4. RESULTS

In the following sections, we report the results for feature selection, machine learning model selection, and the statistical analyses.

4.1 Feature Selection

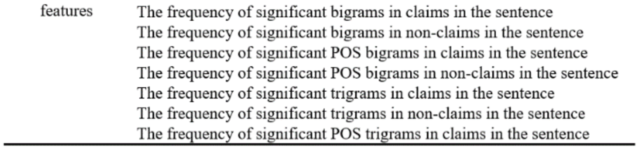
As we have reported in the method section, we applied both content-based features and structure (position) based features to train the model.

Altogether, we had 17 features calculated at the sentence level. Six were structure (position) based features as reported in the method section: (1) the position of the sentence in the whole essay; (2) normalized sentence position in the essay; (3) the position of the paragraph in which the sentence was located; (4) normalized paragraph position in the essay; (5) the position of the sentence in the paragraph where the sentence occurred; and (6) the normalized position in paragraph. Eight of the features were content-based n-gram/POS n-gram frequency calculated based on sentence level. These features included: (1) the frequency of significant bigrams in claims; (2) the frequency of significant bigrams in non-claims; (3) the frequency of significant POS bigrams in claims; (4) the frequency of significant POS bigrams in non-claims; (5) the frequency of significant trigrams in claims; (6) the frequency of significant trigrams in non-claims; (7) the frequency of significant POS trigrams in claims; and (8) the frequency of significant POS trigrams in non-claims. The other three features were word counts, bigram counts, and trigram counts of the sentence.

Before moving forward to build the model, we conducted correlational analyses to remove highly correlated variables. The results of this analysis indicated that the position of the sentence in the essay was highly correlated with normalized sentence position in the essay ($r = .85$, $p < .001$), with the position of paragraph in essay ($r = .91$, $p < .001$), and with normalized paragraph position in essay ($r = .83$, $p < .001$). Normalized sentence position in essay was also highly correlated with the position of the paragraph in essay ($r = .89$, $p < .001$) and normalized paragraph position in essay ($r = .94$, $p < .001$). Meanwhile, the position of paragraph in essay was highly correlated with normalized paragraph position in essay ($r = .92$, $p < .001$). Based on these results, we decided to remove the position of sentence in essay, the paragraph position in essay, and the normalized paragraph position from the independent variables.

For the structure-based features, only the frequency of significant POS trigrams had a strong correlation with the frequency of significant POS bigram ($r = .54$, $p < .001$). For the word and n-gram counts variables, since the variable word counts were highly correlated with bigram counts ($r = 1$, $p < .001$) and trigram counts ($r = 1$, $p < .001$), we decided to remove both of the latter variables and only keep the variable of word counts. After this process, 10 features remained for model development (see Table 2).

Table 2 Summary of structural and content-based features for model development

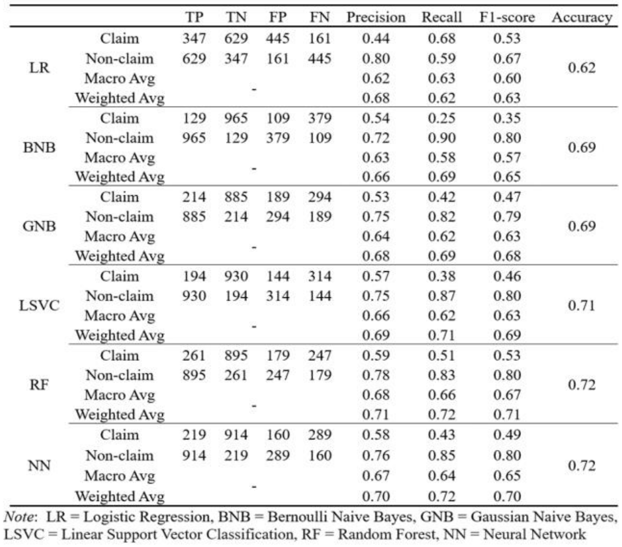


Category	Feature
Structure (positional) features	Normalized sentence position in the essay
	Normalized sentence position in the paragraph
Content-based features	Word counts of the sentence
	The frequency of significant bigrams in claims in the sentence
	The frequency of significant bigrams in non-claims in the sentence
	The frequency of significant POS bigrams in claims in the sentence
	The frequency of significant POS bigrams in non-claims in the sentence
	The frequency of significant trigrams in claims in the sentence
	The frequency of significant trigrams in non-claims in the sentence
	The frequency of significant POS trigrams in claims in the sentence

4.2 Model Selection

We built six different supervised machine learning models on our training data using six different classifiers. We then we used the six models to predict discourse types of sentences in our testing corpus. We evaluated the performance of the models using accuracy, precision, recall, and F1-score. Table 3 reports the performance of the classifiers on claim and non-claim identification in the test set. The Random Forest model was selected as the best model to predict the discourse type in the validation corpus.

Table 3 Performance of the multiple classifiers on claim detection in the test set



		TP	TN	FP	FN	Precision	Recall	F1-score	Accuracy
LR	Claim	347	629	445	161	0.44	0.68	0.53	0.62
	Non-claim	629	347	161	445	0.80	0.59	0.67	
	Macro Avg	-				0.62	0.63	0.60	
	Weighted Avg					0.68	0.62	0.63	
BNB	Claim	129	965	109	379	0.54	0.25	0.35	0.69
	Non-claim	965	129	379	109	0.72	0.90	0.80	
	Macro Avg	-				0.63	0.58	0.57	
	Weighted Avg					0.66	0.69	0.65	
GNB	Claim	214	885	189	294	0.53	0.42	0.47	0.69
	Non-claim	885	214	294	189	0.75	0.82	0.79	
	Macro Avg	-				0.64	0.62	0.63	
	Weighted Avg					0.68	0.69	0.68	
LSVC	Claim	194	930	144	314	0.57	0.38	0.46	0.71
	Non-claim	930	194	314	144	0.75	0.87	0.80	
	Macro Avg	-				0.66	0.62	0.63	
	Weighted Avg					0.69	0.71	0.69	
RF	Claim	261	895	179	247	0.59	0.51	0.53	0.72
	Non-claim	895	261	247	179	0.78	0.83	0.80	
	Macro Avg	-				0.68	0.66	0.67	
	Weighted Avg					0.71	0.72	0.71	
NN	Claim	219	914	160	289	0.58	0.43	0.49	0.72
	Non-claim	914	219	289	160	0.76	0.85	0.80	
	Macro Avg	-				0.67	0.64	0.65	
	Weighted Avg					0.70	0.72	0.70	

Note: LR = Logistic Regression, BNB = Bernoulli Naive Bayes, GNB = Gaussian Naive Bayes, LSVC = Linear Support Vector Classification, RF = Random Forest, NN = Neural Network

4.3 Relationship between Essay Quality and Number of Claims

Spearman's correlations were computed among the number, percentage, and the average positionality of claims and non-claims and the human raters' holistic scores for each essay in the validation corpus. We included text length to assess if the raw scores highly correlated with the number of words in the essay (a strong predictor of essay quality). Correlational analysis indicated the number of predicted claims (r = .35, p < .001) and the average position of non-claims in text (r = -.19, p < .001) showed at least a small effect size (r > .099) with essay quality and were not strongly correlated with text length (r < .70). These variables were selected for inclusion in our regression analysis to predict essay quality scores. However, the percentage of predicted claims (r = .08, p = .015) and non-claims (r = -.08, p = .015) and the average position of claims (r = .04, p < .001) had weak correlations with essay quality.

A significant regression equation was reported (R^2 = .132, F(2,2266) = 172.3, p < .001). The model explained 13.2% of the variance of the human scores. Two significant predictors of essay quality were included in the model: number of claims (β = .132, p < .001) and the average position of non-claims in text (β = -2.829, p < .001).

5. CONCLUSION AND FUTURE WORK

In this study, we extracted content-based linguistic features and structure-based features to train and predict discourse types of claim and non-claim in argumentative essays. The average testing accuracy (F1) of the classifiers used in this study (Logistic Regression, Bernoulli Naive Bayes, Gaussian Naive Bayes, Linear Support Vector Classification, Random Forest, and Neural Network) was around .69. This aligns with the accuracies reported in Stab and Gurevych (2017) to a degree. In their work, they reported F1 scores from an SVM classifier for major claims, claims, and premises using structural, lexical, contextual, syntactic, discourse markers, and embeddings features. Their F1 scores for these features in tandem were .77. F1 scores in isolation were .59 for lexical features, .60 for contextual features, .39 for syntactic features, .52 for discourse features, and .75 for structural features. These results seem to indicate that the individual content-based features (lexical, syntactic, indicator, and contextual features) might have encountered an upper limit in terms of the accuracy of identification if other features were not combined. The accuracy of the identification of claims in our study also seems to support this interpretation.

In terms of application, we found that the number of claims and the average position of non-claims in text were indicators of essay quality. A significant regression model was found to predict holistic human scores based on these variables. The model explained 13% of the variance in the human scores.

To improve the accuracy of classification, we are planning to implement a classifier with more diverse features from a contextual and discourse perspective including contextual, discourse, syntactic, and lexical features. We presume this will increase accuracy based on findings from Stab and Gurevych (2017) who showed that the combination of all features increased their accuracy. We also intend to investigate the relationship between argumentation elements from a broader view by including more argumentation elements such as major claims, primary claims, counterarguments, rebuttals, and conclusions. Further, we plan on annotating the relationships between these discourse elements and build models to automatically identify the discourse elements as well as their functional relationships.

In the current study, we have demonstrated the usefulness of content and structural features in automated claim detection and explored the relations between the number and positionality of claims and writing quality. Our findings can positively supplement existing automated essay scoring (AES) and automated writing evaluation (AWE) systems and may provide implications for the teaching of argumentative essays.

6. ACKNOWLEDGMENTS

This research was supported in part by the Institute for Education Sciences (IES R305A180261), Office of Naval Research N00014-17-1-2300, the Bill & Melinda Gates Foundation, The Chan Zuckerberg Initiative, and Schmidt Futures. Ideas expressed in this material are those of the authors and do not necessarily reflect the views of our funders.

7. REFERENCES

[1] Biber, D., Conrad, S. and Cortes, V., 2004. If you look at…: Lexical bundles in university teaching and textbooks. *Applied linguistics*, *25*(3), pp.371-405.

[2] Bird, S., Klein, E. and Loper, E., 2009. Nltk book.

[3] Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., Lu, C., Nolan, J., Rock, D. and Wolff, S., 1998. Computer analysis of essay content for automated score prediction: A prototype automated scoring system for GMAT analytical writing assessment essays. *ETS Research Report Series*, *1998*(1), pp.i-67.

[4] Burstein, J., Kukich, K., Wolff, S., Lu, C. and Chodorow, M., 2001. Enriching Automated Essay Scoring Using Discourse Marking.

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

[6] Burstein, J., Marcu, D., Andreyev, S. and Chodorow, M., 2001, July. Towards automatic classification of discourse elements in essays. In *Proceedings of the 39th annual Meeting on Association for Computational Linguistics* (pp. 98-105). Association for Computational Linguistics.

[7] Kilgarriff, A., 2001. Comparing corpora. *International journal of corpus linguistics*, *6*(1), pp.97-133.

[8] Klebanov, B.B., Stab, C., Burstein, J., Song, Y., Gyawali, B. and Gurevych, I., 2016, August. Argumentation: Content, structure, and relationship with essay quality. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)* (pp. 70-75).

[9] Lawrence, J. and Reed, C., 2015, June. Combining argument mining techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining* (pp. 127-136).

[10] Marcus, M., Santorini, B. and Marcinkiewicz, M.A., 1993. Building a large annotated corpus of English: The Penn Treebank.

[11] McNamara, D.S., Raine, R., Roscoe, R., Crossley, S.A., Jackson, G.T., Dai, J., Cai, Z., Renner, A., Brandon, R., Weston, J.L. and Dempsey, K., 2012. The Writing-Pal: Natural language algorithms to support intelligent tutoring on writing strategies. In *Applied natural language processing: Identification, investigation and resolution* (pp. 298-311). IGI Global.

[12] Nguyen, H. and Litman, D., 2015, June. Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining* (pp. 22-28).

[13] Nguyen, H. and Litman, D., 2016, August. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1127-1137).

[14] Persing, I. and Ng, V., 2015, July. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 543-552).

[15] Rayson, P. and Garside, R., 2000, October. Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing corpora-Volume 9* (pp. 1-6). Association for Computational Linguistics.

[16] Stab, C. and Gurevych, I., 2014, October. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 46-56).

[17] Stab, C. and Gurevych, I., 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, *43*(3), pp.619-659.

[18] Yumusak, S., Dogdu, E. and Kodaz, H., 2014. Tagging accuracy analysis on part-of-speech taggers. *Journal of Computer and Communications*, *2*(4), pp.157-162.