Automated Claim Identification Using NLP Features in Student Argumentative Essays

Qian Wan
Georgia State University
qwan1@gsu.edu

Scott Crossley
Georgia State University
scrossley@gsu.edu

Michelle Banawan
Arizona State University
mbanawan@asu.edu

Renu Balyan
SUNY at Old Westbury
balyanr@oldwestbury.edu

Yu Tian
Georgia State University
ytian9@gsu.edu

Danielle McNamara
Arizona State University
dsmcnama@asu.edu

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

ABSTRACT

The current study explores the ability to predict argumentative claims in structurally-annotated student essays to gain insights into the role of argumentation *structure* in the quality of persuasive writing. Our annotation scheme specified six types of argumentative components based on the well-established Toulmin's model of argumentation. We developed feature sets consisting of word count, frequency data of key n-grams, positionality data, and other lexical, syntactic, semantic features based on both sentential and suprasentential levels. The suprasentential Random Forest model based on frequency and positionality features yielded the best results, reporting an accuracy of 0.87 and kappa of 0.73. This model will be included in an online writing assessment tool to generate feedback for student writers.

Keywords

Argumentation, Claim identification, Argumentative writing

1. INTRODUCTION

Written argumentation has been an important area of study for many years [43, 45]. Recent developments in natural language processing (NLP) have introduced new approaches to automatically detect the discourse structure of argumentative essays [7, 8, 9, 10, 26, 33, 34, 38, 44, 45]. These studies have shown that content (i.e., lexical, syntactic, and semantic) and structural features (i.e., the positionality of tokens, sentences, and paragraphs) are effective in detecting discourse elements.

Researchers have used fixed discourse markers at the word and phrase levels [5, 12, 18, 42] as indicators of different argumentative structures. This approach has been applied in discourse [17, 19, 22] and NLP analyses [7, 8, 9, 47]. These studies generally identify relations between discourse markers and their functions according to the conceptual framework of conjunctive relations [36]. For instance, phrases such as *in summary* and *in conclusion* are associated with the discourse function of 'summarizing' an argument. Such discourse markers have been used to identify the attributes of the structural elements in argumentative essays [8, 9, 36, 37]. For example, Burstein et al. [7] annotated structural information of argumentative essays collected from TOEFL, GRE, and GMAT. Discourse markers indicating each of the argumentative functions were extracted automatically from the essays. A word list that contained the discourse markers and their corresponding argumentative functions was formed and used to automatically predict instances of argumentation. Similarly, Palau and Moens [37] implemented a context-free ruled-based approach for argumentation mining in legal texts. They focused on and developed rules based on common expressions encountered in the legal documents such as *for these reasons*, *in light of all the material*, and discourse markers, such as *however* or *furthermore*. Using this approach, they obtained accuracy of approximately 0.6 in detecting the argumentation structures, while maintaining F1-measure of around 0.7 for recognizing premises and conclusions in legal texts.

In more recent work, Stab and Gurevych [44, 45] provided publicly available corpora comprising students' argumentative essays and annotation guidelines for parsing argumentations. In these corpora, the essays were annotated based on three major argumentative categories: major claim, claim, and premise. They then used lexical, structural, syntactic, discourse markers, and other features to identify argument components. The lexical features consisted of binary lemmatized unigrams and the 2,000 most frequent bigrams extracted from a training corpus. The structural features captured the position of components in the text and the number of tokens in those components. Discourse markers included logical connectives such as *therefore, thus,* or *consequently* and the use of first-person pronouns (which indicated major claims). The syntactic features included part of speech (POS) distributions, number of sub-clauses, and the tense of the main verb. Using support vector machine models, Stab and Gurevych [45] found that a combination of all these features yielded an F1 score of 0.77. Khatib et al. in [3] employed a classifier for argumentativeness based on the research in [37, 44, 45], and evaluated its performance on student essays from [44]. Khatib et al. used n-grams, syntax, discourse makers and part of speech (POS) features in an argument. Their results indicated that a combination of n-grams, POS tags, and syntax features yielded accuracy of 0.64, 0.62, and 0.59 on classifying arguments in students' essays, while the full feature set model yielded an accuracy of 0.67. Though only unigram through tri-grams were included in the POS feature.

Though the use of discourse markers, n-grams and POS as indicators has been common in the detection of argumentative elements, few studies have examined whether using longer sequences of n-grams (beyond tri-grams) and their POS tags would contribute to identifying argumentative features. We also note that other types of linguistic features related to lexical, structural, cohesion, and affective features were not tested in previous studies [e.g., 37, 45]. Therefore, this study explores a wider range of NLP features, and examines their contribution to model accuracy. We do so specifically on a corpus of student essays annotated on theoretically-aligned classifications of argumentative elements expected in academic settings. This is in contrast to most of the existing corpora in English that are annotated for argumentative structures and are from the domains of law [e.g., 4, 37], biology and medicine [e.g., 20], and user-generated content, e.g., Wikipedia articles or debate data, see [1, 2, 27, 41]. Few corpora [44, 45] have been developed for argumentation mining in the educational settings. In this study, we build on Stab and Gurevych's work [44, 45] by developing a structurally annotated corpus based on the Toulmin model [46] of argumentation that better reflects the structure of student essays. Our objective is for the corpus – and the models of argumentation developed from the corpus – to contribute to the development of writing assessment tools that can deliver useful feedback to student writers.

Thus, in this study, we introduce a new corpus of essays annotated for argumentative features. We then develop NLP approaches to automatically identify claims in structurally annotated argumentative essays using length, frequency data of significant n-grams and POS tags, positionality data, and a wide range of lexical, syntactic, cohesion, and cognitive features extracted from a number of NLP tools [14, 15, 25, 24]. We compared the identification accuracy of multiple machine learning classifiers using different types of derived features at different levels (based on sentences or argumentative elements that are suprasentential). Our goal is to better understand whether and how the selection of the linguistic features, the level of units for identification (both sentential and suprasentential), and the choice of classifiers influence the accuracy of claim identification. Finally, we conduct an error analysis of the best model and discuss the distribution of the misclassification instances and related features. This study is guided by the following research question:

To what extent do length, frequency of significant n-grams (and POS tags of n-grams), lexical, syntactic, and semantic features, and positionality predict argumentative claims in essays?

2. METHOD

2.1 Corpus

For the analysis, we annotated 314 persuasive essays. The essays were written by undergraduate students (*N* = 314) at a public university in the United States who were native speakers of English. Two prompts from retired test banks of the Scholastic Assessment Test (SAT) were used. The prompts were counterbalanced such that half of the students wrote about 'originality and uniqueness' while the other half wrote about 'heroes versus celebrities.' All essays had been scored previously by expert raters for holistic writing quality. For each essay, we extracted the average number of letters per word, the number of words, number of types, type-token ratio, average number of words per sentence, the number of sentences and paragraphs. Descriptive statistics for these items of the 314 essays are reported in Table 1.

Table 1. Descriptive statistics of the persuasive essays

	Mean	SD	Median	Range
Letters per word	4.52	0.24	4.51	1.50
Number of words	354.46	118.20	344.00	680.00
Number of types	178.17	50.01	173.00	279.00
Type-token ratio	0.52	0.07	0.52	0.41
Words per sentence	17.74	4.30	17.06	35.08
Number of sentences	20.65	7.42	20.00	48.00
Number of paragraphs	3.86	1.38	4.00	7.00

2.2 Annotation of argumentative elements

The essays were structurally annotated by normed raters for argumentative elements. We used the modified Toulmin models [46] presented in [35] and [30] as the basis for the annotation rubric. The rubric adopted six elements (i.e., micro-categories) as the building blocks of the argumentation framework: Final Claim, Primary Claim, Counterclaim, Rebuttal, Data, and Concluding Summary. The definitions of each of these elements are presented in Appendix A.

The essays were coded by two annotators on the web-based text annotation platform 'Tagtog'[1]. The two annotators were both native speakers of English and were undergraduate students majoring in applied linguistics at a public university in the United States. Before independent annotation, a norming process was conducted to help ensure consistency in annotations. Once normed, the two annotators worked independently and coded the 314 essays in the opposite order to avoid recency effects.

The two annotators made decisions on both the boundary of an argumentative element and the category of the element. An argumentative element was inherently suprasentential (i.e., according to the annotation scheme derived from the norming session, it could contain one or more sentences, and the content could be over the span of paragraphs). Inter-rater reliability calculated using Fleiss's Kappa for all the annotations was 0.584 ($p < 0.001$), indicating fair to good agreement [16]. Disagreements of either boundary or category of the argumentative elements between the two annotators were adjudicated by an expert adjudicator who had years of experience teaching and conducting writing research. In the case of disagreement, the expert adjudicator compared the annotations from both annotators and made the final decision for both the boundary and the category of the argumentative element.

The current study focuses on the identification of claims versus non-claims, mainly because of the small sample size of the corpus and the distribution of micro-categories. Thus, we combined the categories of Final Claim, Primary Claim, Counterclaim, and Rebuttal into a single category of claims. The remaining categories of Data and Concluding summary were classified as non-claims as was any non-annotated text.

[1] https://www.tagtog.net

2.3 Training and test sets

Annotation of the data led to the classification of 2264 argumentative elements. As mentioned in Section 2.2, the argumentative elements were inherently suprasentential. We further split the elements into sentences to determine whether this influenced accuracy. All sentences from the same argumentative element were given the same annotation as the original category (i.e., claims or non-claims). We thus had two data sets: 1) a sentence-tokenized data set (N = 6326) and 2) a suprasentential data set (N = 2264). We randomly selected 70% of the argumentative elements as the training set, and the remaining 30% of the elements as the test set for both datasets. We report the number of argumentative elements, and number of claims and non-claims for the datasets in Table 2.

Table 2. Numbers of elements, claims and non-claims for the training and test sets

Data set	Number of elements	Number of claims	Number of non-claims
Suprasentential training set	1594	639	955
Suprasentential test set	670	267	403
Sentential training set	4401	935	3466
Sentential test set	1925	409	1516

2.4 Features

2.4.1 Word count

We extracted the number of words for each claim and non-claim at the sentential and suprasentential level.

2.4.2 N-gram frequency

We extracted n-grams and the POS combinations of these n-grams for both claims and non-claims. We assume that some n-grams (or POS n-grams) are more likely to identify claims versus non-claims (and vice versa), and the frequency of these key n-grams (or POS n-grams) could serve as good indicator of the type of an argumentative element or sentence. We used keyness values [21] as the measurement of importance of the n-grams or POS n-grams in claims and non-claims. Keyness values can provide evidence of whether n-grams and POS n-grams are more common in one corpus as compared with the other corpus. In the current study, we treated the claims and non-claims as two separate corpora.

Raw and normalized frequency (i.e., normalized by the total number of words in all claims and non-claims, respectively) for each n-gram (or POS n-gram) that occurred both in claims and non-claims were calculated. The keyness value of each n-gram was also calculated based on the frequency data following Rayson and Garside's guidelines [40]. Specifically, if an n-gram or POS n-gram had a keyness value greater than 3.84 (equivalent to $p < 0.05$), and if it had a higher normalized frequency in claims, it was considered more likely to occur in claims over non-claims, and vice versa. The range of the n-grams and POS n-grams was from unigram to seven-grams. NLTK [6] was used to tokenize the texts into n-grams and label the POS for the n-grams. For example, the following phrases *should be*, *would be*, *can be*, and *will be* were converted to the same POS n-gram combination: MD (modal) + VB (verb base). We did not remove stopwords before n-gram tokenization. For each suprasentential and sentential argumentative element in the training and test sets, we calculated the frequency of each type of the significant n-grams or POS n-grams (e.g., bigrams that were significant in claims), and normalized the frequency by the length (word counts).

2.4.3 Positionality of the elements

Beyond n-gram frequency, studies have shown that, the position of argumentative elements is an indicator of their structural function [e.g., 7, 8, 10]. In this study, two types of normalized positional variables for each argumentative element or sentence were calculated as positionality features.

Normalized element or sentence position in an essay was computed as the ratio of the element/sentence position in an essay to the number of elements/sentences in the essay (e.g., if an argumentative element or a sentence was the 5th element or sentence in an essay of 10 elements/sentences in total, the value of this variable would be 5 divided by 10, or 50%). The normalized position of the element or sentence in a paragraph was computed as the ratio of the element/sentence position in a paragraph to the total number of elements/sentences in that paragraph. That means, if an argumentative element or a sentence was the 2nd element (sentence) in a paragraph, in which there were 5 elements (sentences) in total, the value would be 2 divided by 5, or 40%).

2.4.4 Other lexical, syntactic, and semantic features

To explore whether additional lexical, syntactic, cohesion, and cognitive text features increased the accuracy in identifying claims and non-claims, we extracted 925 features for each of the argumentative elements. These features were extracted using the Suite of Automatic Linguistic Analysis Tools (SALAT) [14, 15, 25, 24]. SALAT includes multiple NLP tools including TAACO (Tool for the Automatic Analysis of Cohesion), TAALES (Tool for the Automatic Analysis of lexical Sophistication), TAASSC (Tool for the Automatic Analysis of Syntactic Sophistication and Complexity), and SEANCE (Sentiment Analysis and Cognition Engine). Two-sample t-tests or Wilcoxon's tests were conducted using the variables after removing SALAT variables that were not normally distributed. We then removed those variables where the results of t-test or Wilcoxon's test were not significant between the group of claims and non-claims. Finally, by visual inspection, 20 out of 131 variables that were relevant to argumentative elements were selected. Hand selection of variables was done to avoid problems of overfitting. The selected NLP features and their descriptions are presented in Appendix B.

2.4.5 Feature reduction

To avoid multicollinearity, we conducted correlation analyses among all the derived features (one versus all) for the two training sets, respectively. If two or more variables correlated with $r >$ 0.699, the variable(s) with the lower correlation with the category of the argumentative element/sentence were removed, and the variable with the higher correlation was retained. The feature reduction process was done on the two training sets first and then applied to the test sets. After feature reduction, the frequency features that were retained included word count (of the argumentative element or sentence), the frequency of the significant unigram in claims and in non-claims, bigrams and quad-grams in claims, and the frequency of significant POS unigrams, trigrams, four-grams, five-grams in claims and in non-claims, and frequency of significant six-grams in claims. The two positionality features and the selected 20 SALAT features were also retained.

Table 3. Model accuracy results

Classifier	Model	Accuracy	Kappa	Label	Precision	Recall	F1
Logistic Regression	Suprasentential - Frequency and positionality	0.852	0.691	Non-Claim	0.874	0.881	0.878
				Claim	0.818	0.809	0.814
	Suprasentential - Full features	0.845	0.675	Non-Claim	0.867	0.876	0.872
				Claim	0.810	0.798	0.804
	Sentential - Frequency and positionality	0.802	0.216	Non-Claim	0.817	0.965	0.885
				Claim	0.604	0.198	0.298
	Sentential - Full features	0.800	0.244	Non-Claim	0.823	0.951	0.882
				Claim	0.569	0.242	0.340
Naive Bayes	Suprasentential - Frequency and positionality	0.769	0.485	Non-Claim	0.747	0.931	0.829
				Claim	0.833	0.524	0.644
	Suprasentential - Full features	0.819	0.618	Non-Claim	0.831	0.878	0.854
				Claim	0.799	0.730	0.763
	Sentential - Frequency and positionality	0.791	0.267	Non-Claim	0.834	0.925	0.878
				Claim	0.515	0.301	0.380
	Sentential - Full features	0.789	0.271	Non-Claim	0.833	0.916	0.872
				Claim	0.506	0.318	0.390
K-Nearest Neighbors	Suprasentential - Frequency and positionality	0.836	0.650	Non-Claim	0.835	0.906	0.869
				Claim	0.837	0.730	0.780
	Suprasentential - Full features	0.787	0.526	Non-Claim	0.760	0.943	0.842
				Claim	0.865	0.551	0.673
	Sentential - Frequency and positionality	0.818	0.286	Non-Claim	0.827	0.973	0.894
				Claim	0.709	0.245	0.364
	Sentential - Full features	0.804	0.196	Non-Claim	0.813	0.976	0.887
				Claim	0.654	0.166	0.265
Support Vector Machines	Suprasentential - Frequency and positionality	0.863	0.714	Non-Claim	0.886	0.886	0.886
				Claim	0.828	0.828	0.828
	Suprasentential - Full features	0.833	0.652	Non-Claim	0.865	0.856	0.860
				Claim	0.786	0.798	0.792
	Sentential - Frequency and positionality	0.818	0.336	Non-Claim	0.839	0.951	0.891
				Claim	0.639	0.325	0.431
	Sentential - Full features	0.822	0.320	Non-Claim	0.833	0.968	0.896
				Claim	0.706	0.281	0.402
Random Forest	Suprasentential - Frequency and positionality	0.873	0.734	Non-Claim	0.886	0.906	0.896
				Claim	0.853	0.824	0.838
	Suprasentential - Full features	0.866	0.720	Non-Claim	0.890	0.886	0.888
				Claim	0.829	0.835	0.832
	Sentential - Frequency and positionality	0.832	0.419	Non-Claim	0.858	0.943	0.898
				Claim	0.664	0.421	0.515
	Sentential - Full features	0.829	0.390	Non-Claim	0.850	0.951	0.897
				Claim	0.672	0.377	0.483

To examine whether adding the SALAT features improved the accuracy of claim identification, we created two versions of the feature sets. The first version comprised the n-gram frequency (including word count) features and positionality features, and the second version comprised all the features (including the SALAT NLP features). Combined with the different levels of discourse units (sentential and suprasentential), four pairs of datasets (training and test sets) were prepared for modeling: the frequency and positionality versions along with the full feature versions at both the sentential and suprasentential levels.

2.4.6 Classifiers

We used the ‘caret’ [23], ‘randomForest’ [28], ‘e1071’ [32], and ‘tidyverse’ packages [48] in R [13] to apply Logistic Regression, Naïve Bayes, K-Nearest Neighbors, Support Vector Machines, and Random Forest models. 10-fold cross validation with five repeats was used. We trained and tested the four versions of data separately. For the SVM classifier, a linear, polynomial, and radial kernel was applied. The model with the best performance was selected to make predictions on the test set.

3. RESULTS

3.1 Model evaluation

The classification performances (precision, recall, F1 scores, accuracy, and Cohen’s kappa) of the multiple models on the test sets are reported in Table 3.

Overall, the models developed on frequency and positionality features slightly outperformed the models developed using all the features. This indicates that adding lexical, syntactic, cohesion, and cognitive NLP features does not improve the accuracy of the classification of claims and non-claims. In terms of the selection of the unit of classification, the suprasentential models outperformed the sentential models. Finally, the suprasentential Random Forest model based on frequency and positionality features yielded the best accuracy (0.873) and Kappa (0.734), followed by the suprasentential model based on the full feature set, which yielded an accuracy of 0.866 and Kappa of 0.720, which represents good performance based on the scale of Cohen's Kappa values [11].

3.2 Important variables

Variable importance for the best model (the suprasentential Random Forest model based on word count, n-gram frequency and positionality features) was reported by the 'caret' package. Table 4 shows the top 10 important variables and their importance values for this model.

The variable importance values showed that the length (word count) of an argumentative element, the normalized position of the argumentative element in the essay, and the frequency of significant bigrams in claims in the argumentative element are the three most important variables.

Table 4. Variable importance values

Variable	Importance Value
Word Count	289.988
Normalized element position in the essay	162.083
Frequency of significant bigrams in claims	47.992
Frequency of significant unigrams in claims	31.147
Normalized element position in the paragraph	29.791
Frequency of significant POS five grams in claims	28.465
Frequency of significant POS four grams in claims	27.389
Frequency of significant unigrams in non-claims	25.399
Frequency of significant POS unigrams in claims	25.272
Frequency of significant POS unigrams in non-claims	23.812
Frequency of significant POS trigrams in claims	20.364
Frequency of significant POS trigrams in non-claims	18.375
Frequency of significant POS four grams in non-claims	13.745
Frequency of significant four grams in claims	8.676
Frequency of significant POS six grams in claims	8.490
Frequency of significant POS five grams in non-claims	4.210

4. ERROR ANALYSES AND DISCUSSION

We conducted error analyses for the two Random Forest suprasentential models (i.e., the models based on the frequency and positionality feature set and the full feature set). Our goal was to examine the misclassifications of the models to better understand elements that may contribute to model accuracy.

We first examined classification rates. Among all incorrectly classified instances, we found more cases in which a claim was misclassified as a non-claim, whereas non-claims were less frequently misclassified as claims. For both models, around 17% of claims were misclassified and non-claims, and around 10% of non-claims were misclassified as claims. These results indicate that, the models are better at identifying non-claims than claims, potentially due to the imbalanced data between the claims and non-claims. Nevertheless, future studies should examine if there are more representative features in claims that can be integrated into our current feature set.

We next examined if essay quality and length influenced the model accuracy. Specifically, for each argumentative element in the two suprasentential test sets, we extracted the following information: holistic score, number of words, number of sentences, and number of paragraphs in the essay where the argumentative element occurred. We examined differences between the argumentative elements that were correctly and incorrectly predicted for these features using t-tests. No differences were reported for essay quality and length in either model. Thus, the classification of argumentative elements was not related to the quality or the length of essays.

We also examined if differences in model accuracy were related to more specific argumentation categories (i.e., micro-categories). As mentioned in Section 2.2, we merged the argumentation categories of Primary Claim, Final Claim, Counterclaim, and Rebuttal from the original annotated corpus into a larger classification of claims (i.e., a macro-classification). We also classified the remaining categories of Data and Concluding Summary along with Non-annotated texts into non-claims. To assess whether the micro-categories influenced classification of the macro-classification, we compared the prediction accuracies among the seven micro-categories.

The results showed that Counterclaims were not misclassified in either model (likely because of their rarity), Concluding Summaries were not misclassified in the frequency- and positionality-based models, but misclassified 3.9% of the time in the full feature model. Data was misclassified around 9% in both models. Meanwhile, the sub-categories that were more frequently misclassified included: Primary Claims (around 14 misclassified), Final Claims (around 21% misclassified), Non-annotated texts (around 22% misclassified), and Rebuttal (2 out of 3, 66.7% misclassified instances in both models). These results were also in line with findings that claims were more frequently misclassified as non-claims.

To further explore what factors affect the misclassifications among the micro-categories of argumentative types, Welch's t-tests were conducted among all NLP features (see Appendix B) used in the full analysis between correct and incorrect classification instances. However, the analysis was done for the sub-category of Counterclaim since all instances under this category were correctly predicted by the two models. Also, we did not conduct t-tests for the micro-category of Rebuttal due to a small sample size (N =3).

Table 5 presents the features for which significant differences were found between the correct and incorrect classification instances in at least two categories of argumentative types. In general, the classification of Primary Claim, Data, Concluding Summary, and Non-annotated texts seemed to be more strongly influenced by linguistic features. Word count was the strongest indicator of misclassification, in which difference were found for each micro-category. The standard deviation of dependents per object of prepositions was another strong predictor of misclassification, which reflects the development of syntactic complexity [25].

Table 5. Features with significant differences between correct and incorrect classification instances

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Primary claim		Yes	Yes		Yes	Yes	Yes			Yes		Yes	Yes	
Final claim		Yes									Yes			Yes
Data	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes			Yes	Yes	Yes	
Concluding summary	Yes	Yes						Yes	Yes		Yes		Yes	Yes
Nonannotated	Yes	Yes		Yes					Yes	Yes		Yes	Yes	

Note. Shaded gray cells with 'Yes' indicate significant difference ($p < .05$) were found between the correct and incorrect instances. 1 = Number of named entities, 2 = Word count, 3 = Normalized element position in the paragraph, 4 = Normalized element position in the essay, 5 = Frequency of significant unigrams in claims, 6 = Frequency of significant POS trigrams in claims, 7 = Frequency of significant quad-grams in claims, 8 = Hu Liu proportion score, 9 = Objects component score, 10 = Brown frequency score, 11 = Bigram lemma type-token ratio, 12 = Nouns as modifiers score, 13 = Dependents per object of the preposition (SD), 14 = T-units per sentence.

The number of named entities was a strong indicator for the non-claims, wherein the incorrect instances of non-claims contained fewer named entities versus the correct instances. The nouns as modifier scores were also predictive of misclassification, which measured the use of nouns as nominal modifiers in general and the variation in the number of modifiers per nominal [25]. Other linguistics features that influenced the classification accuracy included: the normalized position of the element in paragraph and in essay, the bigram type-token ratio, the frequency of key unigram, quad-gram, and POS trigram in claims, the number of T-units per sentence, the number terms that reference objects, the proportion of the number of words with positive sentiments to the words with negative sentiments, and the mean frequency score based on London-Lund Corpus of Conversation.

5. CONCLUSION

In this study, we proposed an approach that combined the frequency, positionality, and other lexical, syntactic, cohesion, and cognitive NLP features to predict claims and non-claims in argumentative essays. Our model performed well in the classification of these argumentative elements. Our exploration of the features, the comparison between sentential versus suprasentential models, and investigation of the factors that influenced classification accuracy in the error analyses should contribute to the field of automated identification and evaluation of discourse elements in argumentative writing.

It is important to note that the corpus used for this study was relatively small, comprising 314 student essays. Thus, to gain higher accuracies and reliabilities in classifying argumentative elements, we plan on annotating more essays and expanding the current corpus. That also means we will use essays written to more prompts allowing us to extract key n-grams and POS n-grams that are more generic and less restricted to the specific prompts used here. In addition, due to the small sample size, our classification of argumentative elements was simplified to focus on claims versus non-claims. We are interested in exploring the classification of the micro-categories (Primary Claim, Final Claim, Counter Claim, Rebuttal, Data, and Concluding Summary) in a larger corpus. We also plan to include the prediction of the quality of these argumentative elements in students' writing.

The models developed in this study will be included in an online Writing Assessment Tool (WAT). Implementing the classification algorithm within WAT, WAT's automatic writing evaluation (AWE) system will have the capacity to predict the number of claims in the essay and whether the claims mention the key n-grams that reflects the argument topic. This will afford providing feedback to students on argumentation quality within student essays. The study also provides insight into the length, position, content (e.g., the key n-grams), and other NLP features in claims versus non-claims in students' writing, which will contribute to finer-grained feedback components in our AWE system.

This study also provides important information for others who are developing AWE algorithms to drive feedback on argumentative essays, or more broadly to better understand the use of claims in essays. Specifically, the results of this study inform features related to feedback that can be provided to students about the number of claims, mentioning the argument topic, how to better position argumentative elements within their essays, and how to pay attention to specific linguistic features (such as the use of named entities when giving evidence) in their writing. This is an important achievement in the realm of writing feedback given the crucial need to automate feedback to students on their use of claims and evidence in argumentative essays.

Another important contribution of this study is that we also introduce a new corpus of essays annotated for argumentative elements, which is made publicly available at linguisticanalysistools.org. This corpus includes theoretically aligned argumentative elements that complement existing corpora [44, 45] and adds new components including prompts, holistic scores, additional categories of argumentation, and different educational settings. As such, this study provides the opportunity for other scientists to build upon our work such that we can better understand writing, and the features related to successful composition.

6. ACKNOWLEDGMENTS

The research reported here was supported by the Chan Zuckerberg Initiative, the Bill & Melinda Gates Foundation, and Schmidt Futures through grants to Georgia State University. Additional funding was provided by the Institute of Education Sciences, U.S. Department of Education, and the Office of Naval Research, through Grants R305A180261, R305A180144, N00014-20-1-2623, N00014-19-1-2424 to Arizona State University and University of New Hampshire. The opinions expressed are those of the authors and do not represent views of the funding agencies.

7. REFERENCES

[1] Aharoni, E., Polnarov, A., Lavee, T., Hershcovich, D., Levy, R., Rinott, R., Gutfreund, D. and Slonim, N., 2014, June. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the first workshop on argumentation mining* (pp. 64-68).

[2] Ajjour, Y., Alshomary, M., Wachsmuth, H. and Stein, B., 2019, November. Modeling frames in argumentation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 2915-2925).

[3] Al Khatib, K., Wachsmuth, H., Hagen, M., Köhler, J. and Stein, B., 2016, June. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 1395-1404).

[4] Ashley, K.D. and Walker, V.R., 2013, November. From Information Retrieval (IR) to Argument Retrieval (AR) for Legal Cases: Report on a Baseline Study. In *JURIX* (pp. 29-38).

[5] Biber, D. and Conrad, S., 1999. Lexical bundles in conversation and academic prose. *Language and Computers*, *26*, pp.181-190.

[6] Bird, S., Klein, E. and Loper, E., 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

[7] 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.

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

[9] 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.

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

[11] Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), pp.37-46.

[12] Cohen, R., 1984, July. A computational theory of the function of clue words in argument understanding. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics* (pp. 251-258).

[13] Core Team, R., 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing. *Vienna, Austria: URL https://www. R-project. org/.[Google Scholar]*.

[14] Crossley, S.A., Kyle, K. and McNamara, D.S., 2016. The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, *32*, pp.1-16.

[15] Crossley, S.A., Kyle, K. and McNamara, D.S., 2017. Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior research methods*, *49*(3), pp.803-821.

[16] Fleiss, J.L., Levin, B. and Paik, M.C., 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions*, *2*(212-236), pp.22-23.

[17] Fraser, B., 1999. What are discourse markers?. *Journal of pragmatics*, *31*(7), pp.931-952.

[18] Grosz, B. and Sidner, C.L., 1986. Attention, intentions, and the structure of discourse. *Computational linguistics*.

[19] Hirschberg, J. and Litman, D., 1993. Empirical studies on the disambiguation of cue phrases. *Computational linguistics*, *19*(3), pp.501-530.

[20] Houngbo, H. and Mercer, R.E., 2014, June. An automated method to build a corpus of rhetorically-classified sentences in biomedical texts. In *Proceedings of the first workshop on argumentation mining* (pp. 19-23).

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

[22] Knott, A. and Dale, R., 1994. Using linguistic phenomena to motivate a set of coherence relations. *Discourse processes*, *18*(1), pp.35-62.

[23] Kuhn, M., 2015. A Short Introduction to the caret Package. *R Found Stat Comput*, *1*.

[24] Kyle, K. and Crossley, S.A., 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, *49*(4), pp.757-786.

[25] Kyle, K., 2016. Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication.

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

[27] Levy, R., Bilu, Y., Hershcovich, D., Aharoni, E. and Slonim, N., 2014, August. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 1489-1500).

[28] Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. *R news*, *2*(3), pp.18-22.

[29] Lippi, M. and Torroni, P., 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, *16*(2), pp.1-25.

[30] Liu, F. and Stapleton, P., 2014. Counterargumentation and the cultivation of critical thinking in argumentative writing: Investigating washback from a high-stakes test. *System*, *45*, pp.117-128.

[31] Max, K., 2016. Contributions from Jed Wing. *Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team,*

Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang and Can Candan, pp.6-0.

[32] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.C. and Lin, C.C., 2015. Misc functions of the department of statistics, probability theory group (formerly: E1071). *Package e1071. TU Wien*.

[33] 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).

[34] 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).

[35] Nussbaum, E.M., Kardash, C.M. and Graham, S.E., 2005. The Effects of Goal Instructions and Text on the Generation of Counterarguments During Writing. *Journal of Educational Psychology*, *97*(2), p.157.

[36] Ong, N., Litman, D. and Brusilovsky, A., 2014, June. Ontology-based argument mining and automatic essay scoring. In *Proceedings of the First Workshop on Argumentation Mining* (pp. 24-28).

[37] Palau, R.M. and Moens, M.F., 2009, June. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law* (pp. 98-107).

[38] 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).

[39] Quirk, R., 1985. The English language in a global context. *English in the world: Teaching and learning the language and literatures*, *16*, pp.17-21.

[40] Rayson, P. and Garside, R., 2000, October. Comparing corpora using frequency profiling. In *The workshop on comparing corpora* (pp. 1-6).

[41] Rinott, R., Dankin, L., Alzate, C., Khapra, M.M., Aharoni, E. and Slonim, N., 2015, September. Show me your evidence-an automatic method for context dependent evidence detection. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 440-450).

[42] Schiffrin, D., 2001. Discourse markers: Language, meaning, and context. *The handbook of discourse analysis*, *1*, pp.54-75.

[43] Song, Y., Heilman, M., Klebanov, B.B. and Deane, P., 2014, June. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining* (pp. 69-78).

[44] 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).

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

[46] Toulmin, S.E., 2003. *The uses of argument*. Cambridge university press.

[47] Van Eemeren, F.H., Houtlosser, P. and Henkemans, A.F.S., 2008. Dialectical profiles and indicators of argumentative moves. *Journal of Pragmatics*, *40*(3), pp.475-493.

[48] Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D.A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J. and Kuhn, M., 2019. Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), p.1686.

APPENDIX

A. Definitions of argumentative elements

Elements	Definitions	Examples
Final Claim	An opinion or conclusion on the main question	In my opinion, every individual has an obligation to think seriously about important matters, although this might be difficult.
Primary Claim	A claim that supports the final claim.	The next reason why I agree that every individual has an obligation to think seriously about important matters is that this simple task can help each person get ahead in life and be successful.
Counterclaim	A claim that refutes another claim or gives an opposing reason to the final claim.	Some may argue that obligating every individual to think seriously is not necessary and even annoying as some people may choose to just follow the great thinkers of the nation.
Rebuttal	A claim that refutes a counterclaim.	Even though people can follow others' steps without thinking seriously in some situations, the ability to think critically for themselves is a very important survival skill.
Data	Ideas or examples that support primary claims, counterclaims, or rebuttals.	For instance, the presidential debate is currently going on. In order to choose the right candidate, voters need to research all sides of both candidates and think seriously to make a wise decision for the good of the whole nation.
Concluding Summary	A concluding statement that restates the claims.	To sum up, thinking seriously is important in making decisions because each decision has an outcome that affects lives. It is also important because if you think seriously it can help you succeed.
Non-annotated	Any text that doesn't fall into any of the above categories	People always strive to be unique or different. This idea clashes with creativeness all through our lives.

B. Descriptions of the SALAT NLP features

NLP features from SALAT	Descriptions
Bigram lemma type-token ratio	Number of unique bigram lemmas (types) divided by the number of total bigram lemmas (tokens)
Brown frequency score	Mean word frequency score based on London-Lund Corpus of Conversation
Brysabaert concreteness score	Sum of concreteness scores based on all words divided by number of words with concreteness scores
COCA academic bigram association strength	Sum of approximate collexeme strength score divided by the number of bigrams in text with collexeme scores
Dependents per clause (SD)	The standard deviation of the total number of dependents per clause
Dependents per object of the preposition (SD)	This score captures the variation (standard deviation) in the prepositional objects
Direct objects per clause	The number of direct objects per clause
Free association tokens response score	Number of response tokens elicited by word as stimuli in discrete word association experiment (based on function words)
Hu Liu proportion score	Proportion of the number of words with positive sentiments to the words with negative sentiments
LDA age of exposure score	Based on Incremental Age of Exposure for words across 13 grade levels; calculated based on 1/slope of linear regression
Lexical decision time	Standardized lexical decision reaction time across all participants for this word (z-score, based on function words)
Nouns as modifiers score	This score captures the use of nouns as modifiers and modifier variation
Number of named entities	The number of named entities
Number of prepositions per clause	This score captures capture noun phrase elaboration and clause complexity
Objects component score	This component score represents the number of terms that reference objects
Possessives component score	This component score captures the use of possessives in general, and specifically captures the use of possessives in nominal subjects, direct objects, and prepositional objects
Sentiment score of dominance	This score captures the sentiment of dominance, measured by the number of words of dominance
Sentiment score of overstating	This score captures the sentiment of overstating, calculated based on words indicating emphasis in realms of frequency, causality, accuracy, validity...
T-units per sentence	Number of T-units in text divided by number of sentences in text
Verb argument constructions association strength	Average approximate collostructional strength score based on the COCA academic corpus

Note. For more information about the SALAT NLP features, please see https://www.linguisticanalysistools.org/