Modeling Stance in Student Essays

Isaac Persing and **Vincent Ng**
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75083-0688
{persingq,vince}@hlt.utdallas.edu

Abstract

Essay stance classification, the task of determining how much an essay's author agrees with a given proposition, is an important yet under-investigated subtask in understanding an argumentative essay's overall content. We introduce a new corpus of argumentative student essays annotated with stance information and propose a computational model for automatically predicting essay stance. In an evaluation on 826 essays, our approach significantly outperforms four baselines, one of which relies on features previously developed specifically for stance classification in student essays, yielding relative error reductions of at least 11.3% and 5.3%, in micro and macro F-score, respectively.

1 Introduction

State-of-the-art automated essay scoring engines such as E-*rater* (Attali and Burstein, 2006) do not grade essay content, focusing instead on providing diagnostic trait feedback on categories such as grammar, usage, mechanics, style and organization. Hence, persuasiveness and other content-dependent dimensions of argumentative essay quality are largely ignored in existing automated essay scoring research. While full-fledged content-based essay scoring is still beyond the reach of state-of-the-art essay scoring engines, recent work has enabled us to move one step closer to this ambitious goal by analyzing essay content, attempting to determine the argumentative structure of student essays (Stab and Gurevych, 2014) and the persuasiveness of the arguments made in these essays (Persing and Ng, 2015).

Stance classification is an important first step in determining how persuasive an argumentative student essay is because persuasiveness depends on how well the author argues *w.r.t. the stance she takes* using the supporting evidence she provides. For instance, if her stance is *Agree Somewhat*, a persuasive argument would involve explaining what reservations she has about the given proposition. As another example, an argumentative essay in which the author takes a *neutral* stance or the author presents evidence that does not support the stance she claims to take should receive a low persuasiveness score.

Given the important role played by stance classification in determining an essay's persuasiveness, our goal in this paper is to examine stance classification in argumentative student essays. While there is a large body of work on stance classification[1], stance classification in argumentative essays is largely under-investigated and is different from previous work in several respects. First, in automated essay grading, the majority of the essays to be assessed are written by students who are learners of English. Hence our stance classification task could be complicated by the authors' lack of fluency in English. Second, essays are longer and more formally written than the text typically used in previous stance classification research (e.g., debate posts). In particular, a student essay writer typically expresses her stance on the essay's topic in a thesis sentence/clause, while a debate post's author may never even explicitly express her stance. Although the explicit expression of stance in essays seems to make our task easier,

[1]Previous approaches to stance classification have focused on three discussion/debate settings, namely congressional floor debates (Thomas et al., 2006; Bansal et al., 2008; Balahur et al., 2009; Yessenalina et al., 2010; Burfoot et al., 2011), company-internal discussions (Agrawal et al., 2003; Murakami and Raymond, 2010), and online social, political, and ideological debates (Wang and Rosé, 2010; Biran and Rambow, 2011; Walker et al., 2012; Abu-Jbara et al., 2013; Hasan and Ng, 2013; Boltužić and Šnajder, 2014; Sobhani et al., 2015; Sridhar et al., 2015).

Prompt	Prompt Parts
Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value.	1) Most university degrees are theoretical. 2) Most university degrees do not prepare students for the real world. 3) Most university degrees are of very little value.
The prison system is outdated. No civilized society should punish its criminals: it should rehabilitate them.	1) The prison system is outdated. 2) No civilized society should punish its criminals. 3) Civilized societies should rehabilitate criminals.

Table 1: Some examples of essay prompts and their associated parts.

identifying stancetaking text in the midst of non-stancetaking sentences in a potentially long essay, as we will see, is by no means a trivial task.

To our knowledge, the essay stance classification task has only been attempted by Faulkner (2014). However, the version of the task we address is different from his. First, Faulkner only performed two-class stance classification: while his corpus contains essays labeled with *For* (Agree), *Against* (Disagree), and *Neither*, he simplified the task by leaving out the arguably most difficult-to-identify stance, *Neither*. In contrast, we perform *fine-grained* stance classification, where we allow essay stance to take one of six values: *Agree Strongly*, *Agree Somewhat*, *Neutral*, *Disagree Somewhat*, *Disagree Strongly*, and *Never Addressed*, given the practical need to perform fine-grained stance classification in student essays, as discussed above. Second, given that many essay prompts are composed of multiple simpler propositions (e.g., the prompt "Most university degrees are theoretical and do not prepare students for the real world" has two parts, "Most university degrees are theoretical" and "Most university degrees do not prepare students for the real world."), we manually split such prompts into *prompt part*s and determine the stance of the author w.r.t. each part, whereas Faulkner assigned an overall stance to a given prompt regardless of whether it is composed of multiple propositions. The distinction is important because an analysis of our annotations described in Section 2 shows that essay authors take different stances w.r.t. different prompt parts in 49% of essays, and in 39% of essays, authors even take stances with different polarities w.r.t. different prompt parts.

In sum, our contributions in this paper are two-fold. First, we propose a computational model for essay stance classification that outperforms four baselines, including our re-implementation of Faulkner's approach. Second, in order to stimulate further research on this task, we make our annotations publicly available. Since progress on this task is hindered in part by the lack of a publicly annotated corpus, we believe that our data set will be a valuable resource for the NLP community.

2 Corpus

We use as our corpus the 4.5 million word International Corpus of Learner English (ICLE) (Granger et al., 2009), which consists of more than 6000 essays on a variety of different topics written by university undergraduates from 16 countries and 16 native languages who are learners of English as a Foreign Language. 91% of the ICLE texts are written in response to prompts that trigger argumentative essays, and thus are expected to take a stance on some issue. We select 11 such prompts, and from the subset of argumentative essays written in response to them, we select 826 essays to annotate for training and testing our stance classification system.[2] Table 1 shows two of the 11 topics selected for annotation.

We pair each of the 826 essays with each of the prompt parts to which it responds, resulting in 1,593 *instance*s.[3] We then familiarize two human annotators, both of whom are native speakers of English, with the stance definitions in Table 2 and ask them to assign each instance the stance label they believe the essay's author would have chosen if asked how strongly she agrees with the prompt part. We additionally furnish the annotators with descriptions of situations that might cause an author to select the more ambiguous classes. For example, an author might choose Agree Somewhat if she appears to mostly agree with the prompt part, but qualifies her opinion in a way that is not captured by the prompt part's bluntness (e.g. an author who claims the prison system in a lot of countries is outdated would Agree Somewhat with the first part of Table 1's second prompt). Or she may choose Disagree Somewhat if she appears to dis-

[2] See our website at `http://www.hlt.utdallas.edu/~persingq/ICLE/` for the complete list of essay stance annotations.

[3] We do not segment the essays' texts according to which prompt part is being responded to. Each (entire) essay is viewed as a response to all of its associated prompt parts.

Stance	Definition
Agree Strongly (885)	The author seems to agree with and care about the claim.
Agree Somewhat (148)	The author generally agrees with the claim, but might be hesitant to choose "Agree Strongly".
Neutral (28)	The author agrees with the claim as much as s/he disagrees with it.
Disagree Somewhat (91)	The author generally disagrees with the claim, but might be hesitant to choose "Disagree Strongly".
Disagree Strongly (416)	The author seems to disagree with and care about the claim.
Never Addressed (25)	A stance cannot be inferred because the proposition was never addressed.

Table 2: Stance label counts and definitions.

agree with the prompt part, but mentions the disagreement only in passing because she does not care much about the topic.

To ensure consistency in annotation, we randomly select 100 essays (187 instances) for annotation by both annotators. Their labels agree in 84.5% of the instances, yielding a Cohen's (1960) Kappa of 0.76. Each case of disagreement is resolved through discussion between the annotators.

3 Baseline Stance Classification Systems

In this section, we describe four baseline systems.

3.1 Agree Strongly Baseline

Given the imbalanced stance distribution shown in Table 2, we create a simple but by no means weak baseline, which predicts that every instance has most frequent class label (Agree Strongly), regardless of the prompt part or the essay's contents.

3.2 N-Gram Baseline

Previous work on stance classification, which assumes that stance-annotated training data is available for every topic for which stance classification is performed, has shown that the N-Gram baseline is a strong baseline. Not only is this assumption unrealistic in practice, but it has led to undesirable consequences. For instance, the proposition "feminists have done more harm to the cause of women than good" elicits much more disagreement than normal. So, if instances from this proposition appeared in both the training and test sets, the unigram feature "feminist" would be strongly correlated with the disagreement classes even though intuitively it tells us nothing about stance. This partly explains why the N-Gram baseline was strong in previous work (Somasundaran and Wiebe, 2010). In light of this problem, we perform leave-one-out cross validation where we partition the instances by *prompt*, leaving the instances created for one prompt out in each test set.

To understand how strong n-grams are when evaluated in our leave-one-prompt-out cross-validation setting, we employ them as features in our second baseline. Specifically, we train a multiclass classifier on our data set using a feature set composed solely of unigram, bigram, and trigram features, each of which indicates the number of times the corresponding n-gram is present in the associated essay.

3.3 Duplicated Faulkner Baseline

While it is true that no system exists for solving our exact problem, the system proposed by Faulkner (2014) comes fairly close. Hence, as our third baseline, we train a multiclass classifier on our data set for fine-grained essay stance classification using the two types of features proposed by Faulkner, as described below.

Part-of-speech (POS) generalized dependency subtrees. Faulkner first constructs a lexicon of stance words in the style of Somasundaran and Wiebe (2010). The lexicon consists of (1) the set of stemmed first unigrams appearing in all stance-annotated text spans in the Multi-Perspective Question Answering (MPQA) corpus (Wiebe et al., 2005), and (2) the set of boosters (clearly, decidedly), hedges (claim, estimate), and engagement markers (demonstrate, evaluate) from the appendix of Hyland (2005). He then manually removes from this list any words that appear not to be stancetaking, resulting in a 453 word lexicon.

Stance words target propositions, which Faulkner notes, usually contain some opinion-bearing language that can serve as a proxy for the targeted proposition. In order to find the locations in an essay where a stance is being taken, he first finds each stance word in the essay. Then he finds the shortest path from the stance word to an opinion word in the sentence's dependency tree, using the MPQA subjectivity lexicon of opinion words (Wiebe et al., 2005). If this nearest opinion word appears in the stance word's immediate or embedded clause, he creates a binary feature by concatenating all the words in the dependency path, POS generalizing all words other than the stance and opinion word, and finally prepending

"not" if the stance word is adjacent to a negator in the dependency tree. Thus given the sentence "I **can** only say that this statement is completely *true*." he would add the feature *can-V-true*, which suggests agreement with the prompt.

Prompt topic words. Recall that for the previous feature type, a feature was generated whenever an opinion word occurred in a stance word's immediate or embedded clause. Each content word in this clause is used as a binary feature if its similarity with one of the prompt's content words meets an empirically determined threshold.

3.4 N-Gram+Duplicated Faulkner Baseline

To build a stronger baseline, we employ as our fourth baseline a classifier trained on both n-gram features and duplicated Faulkner's features.

4 Our Approach

Our approach to stance classification is a learning-based approach where we train a multiclass classifier using four types of features: n-gram features (Section 3.2), duplicated Faulkner's features (Section 3.3), and two novel types of features, stancetaking path-based features (Section 4.1) and knowledge-based features (Section 4.2).

4.1 Stancetaking Path-Based Features

Recall that, in order to identify his POS generalized dependency subtrees, Faulkner relies on two lexica, a lexicon of stancetaking words and a lexicon of opinion-bearing words. He then extracts a feature any time words from the two lexica are syntactically close enough. A major problem with this approach is that the lexica are so broad that nearly 80% of sentences in our corpus contain text that can be identified as stancetaking using this method. Intuitively, an essay may state its stance w.r.t. a prompt part in a thesis or conclusion sentence, but most of essay's text will be at most tangentially related to any particular prompt part. For this reason, we propose to identifying stancetaking text to target *only text that appears directly related to the prompt part*. Below we first show how we identify and stance-labeling relevant stancetaking dependency paths, and then describe the features we derive from these paths.

4.1.1 Identifying relevant stancetaking paths

As noted above, we first identify stancetaking text that appears directly related to the prompt part.

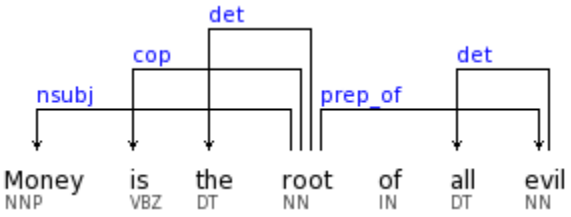

Figure 1: Automatic dependency parse of a prompt part.

To begin, we note that the prompt parts themselves must express a stance on a topic if they can be agreed or disagreed with. By examining the dependency parses[4] of the prompt parts, we can recognize elements of how stancetaking text is structured. From the prompt part shown in Figure 1, for example, we notice that the important words that express a stance in the sentence are "money", "root", and "evil". By analyzing the dependency structure in this and other prompt parts, we discovered that stancetaking text often consists of (1) a **subject** word, which is the child in an nsubj or nsubjpass relation, (2) a **governor** word which is the subject's parent, and (3) an **object**, which is a content word from which there is a (not always direct) dependency path from the governor. We therefore abstract a stance in an essay as a dependency path from a subject to an object that passes through the governor. Thus, the stancetaking dependency path we identify from the prompt part shown in Figure 1 could be represented as money-root-evil.

The obvious problem with identifying stancetaking text in this way is that nearly all sentences contain this kind of stancetaking structure, and just as with Faulkner's dependency paths, there is little reason to believe that any particular path is relevant to an instance's prompt part. Does this mean that nearly all sentences are stancetaking? We would argue that they can be, as even sentences that appear on their face to be mere statements of fact with no apparent value judgment can be viewed as taking a stance on the factuality of the statement, and people often disagree about the factuality of statements. For this reason, after we have identified a stancetaking path, we must determine whether the stance being expressed is relevant to the prompt part before extracting features from it.

[4]Dependency parsing, POS tagging, and lemmatization are performed automatically using the Stanford CoreNLP system (Manning et al., 2014)

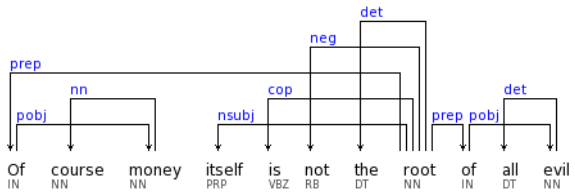

Figure 2: Automatic dependency parse of an essay sentence.

For this reason, we ignore all stancetaking paths that do not meet the following three relevance conditions. First, the lemma of the path's governor must match the lemma of a governor in the prompt part. Second, the lemma of the path's object must match the lemma of some content word[5] in the prompt part. Finally, the containing sentence must not contain a question mark or a quotation, as such sentences are usually rhetorical in nature. We do not require that the subject word match the prompt part's subject word because this substantially reduces coverage for various reasons. For one, of the three words (subject, governor, object), the subject is the word most likely to be replaced with some other word like a pronoun, and possibly because the essays were written by non-native English speakers, automatic coreference resolution cannot reliably identify these cases. We also do not fully trust that the subject identified by the dependency parser will reliably match the subject we are looking for. Given these constraints, we can automatically identify the "itself-root-of-evil" dependency path in Figure 2 as a relevant stancetaking path.

4.1.2 Stance-labeling the paths

Next, we determine whether a stancetaking path identified in the previous step appears to agree or disagree with the prompt part.

To begin, we count the number of negations occurring in the prompt part. Any word like "no", "not", or "none" counts as a negation unless it begins a non-negation phrase like "no doubt" or "not only".[6] Thus, the count of negations in the prompt part in Figure 1 is 0.

After that, we count the number of times the identified stancetaking path is negated. Because these paths occur in student essays and are therefore often not as simply-stated as the prompt parts, this is a little bit more complicated than just counting the containing sentence's negations since the sentence may contain a lot of additional material. To do this, we construct a list of all the dependency nodes in the stancetaking path as well as all of their dependency tree children. We then remove from this list any node that, in the sentence, occurs after the last node in the stancetaking path. The total negation count we are looking for is the number of nodes in this list that correspond to negation words (unless the negation word begins a negation phrase). Thus, because the word "not" is the child of "root" in the path "itself-root-of-evil" we identified in Figure 2, we consider this path to have been negated one time.

Finally, we sum the prompt part negations and the stancetaking path negations. If this sum is even, we believe that the relevant stancetaking path agrees with the prompt part in the instance. If it is odd, however (as in the case of the prompt part and stancetaking text in the dependency tree figures), we believe that it disagrees with the prompt part. To illustrate why we are concerned with whether this sum is even, consider the following examples. If both the prompt part and the stancetaking text are negated, both disagree with the opposite of the prompt part's stance. Thus, they agree with each other, and their negation sum is even (2). If the stancetaking path was negated twice, however, the sum would be odd (3) due to the stance path's double negations canceling each other out, and the stancetaking path would disagree with the prompt part.

4.1.3 Deriving path-based features

We extract four features from the relevant stancetaking dependency paths identified and stance-labeled so far, as described below.

The first feature encodes the count of relevant stancetaking paths that appear to agree with the prompt part. The second feature encodes the count of relevant stancetaking paths that appear to disagree with the prompt part. While we expect these first two features to be correlated with the agreement and disagreement classes, respectively, they may not be sufficient to distinguish between agreeing and disagreeing instances. It is possible, for example, that both features may be greater than zero in a single instance if we have identified one stancetaking path that appears to agree

[5]For our purpose, a content word (1) is a noun, pronoun, verb, adjective, or adverb, (2) is not a stopword, and (3) is at the root, is a child in a dobj or pobj relation, or is the child in a conj relation whose parent is the child in a dobj or pobj relation in the dependency tree.

[6]See our website at `http://www.hlt.utdallas.edu/~persingq/ICLE/` for our list of manually constructed negation words and non-negation phrases.

with the prompt part and another stancetaking path that appears to disagree with the prompt part. It is not clear whether this situation is indicative of only the Neutral class, or perhaps it indicates partial (Somewhat) (Dis)Agreement, or maybe our method of detecting disagreement is not reliable enough, and it therefore makes sense, when we get these conflicting signals, to ignore them entirely and just assign the instance to the most frequent (Agree Strongly) class. For that matter, if neither feature is greater than zero, does this mean that the instance Never Addressed the prompt part, or does it instead mean that our method for identifying stancetaking paths doesn't have high enough recall to work on all instances? We let our learner sort these problems out by adding two more binary features to our instances, one which indicates that both of the first two features are zero, and one that indicates whether both are greater than zero.

4.2 Knowledge-Based Features

Our second feature type is composed of five linguistically informed binary features that correspond to five of the six classes in our fine-grained stance classification task. Intuitively, if an instance has one of these features turned on, it should be assigned to the feature's corresponding class.

1. Neutral. Stancetaking text indicating neutrality tends to be phrased somewhat differently than stancetaking text indicating any other class. In particular, neutral text often makes claims that are about the prompt part's subject, but which are tangential to the proposition expressed in the prompt part. For this reason, we search the essay for words that match the prompt part's subject lemmatically.

After identifying a sentence that is about the prompt part's subject in this way, we check whether the sentence begins with any neutral indicating phrase.[7] If we find a sentence that both begins with a neutral phrase and is about the prompt part's subject, we turn the Neutral feature on. Thus, sentences like the following can be captured: "*In all probability* university students wonder whether or not they spend their time uselessly in studying through four or five years in order to take their *degree*."

2. (Dis)Agree Somewhat. In order to set the values of the features associated with the Somewhat classes, we first identify relevant stancetaking paths as described above. We then trim the list of paths by removing any path whose governor or subject does not have a hedge word as an adverb modifier child in the dependency tree.[8] Thus, we are able to determine that the essay containing the sentence "There is *nearly* no place left for dream and imagination" is likely to belong to one of the Somewhat classes w.r.t. the prompt part "There is no longer a place for dreaming and imagination."

The question now is how to determine which (if any) of the Somewhat classes it should belong to. We analyze all the paths from the list for negation in much the same way we described above, but with one major difference. We hypothesize that when taking a Somewhat stance, students are more likely to explicitly state that the stance being taken is their opinion rather than stating the stance bluntly without attribution. For example, one Disagree Somewhat essay includes the sentence, "*I* **never** *believed* these people were honest if saying that money is just the root of all evil." In order to determine that this sentence contains an indication of the Disagree Somewhat class, we need to account for the negation that occurs at the beginning, far away from the stancetaking path (money-root-of-evil). To do this, we semantically parse the sentence using SEMAFOR (Das et al., 2010). Each of the *semantic frames* detected by SEMAFOR describes an event that occurs in a sentence, and the event's *frame elements* may be the people or other entities that participate in the event. One of the semantic frames detected in this example sentence describes a Believer (I) and the content of his or her belief (all the text after "believed"). Because the sentence includes a semantic frame that (1) contains a first person (I, we) Cognizer, Speaker, Perceiver, or Believer element, (2) contains an element that covers all the text in the dependency path (a Content frame element, in this case), and (3) the word that triggers the frame ("believed") has a negator child in the dependency tree, we add one to this relevant stancetaking path's negation count. This makes this hedged stancetaking path's negation count odd, so we believe that this sentence likely disagrees with its instance's prompt part somewhat. If we find a hedged stancetaking

[7] We construct a list of neutral phrases for introducing another person's ideas from a writing skills website (`http://www.myenglishteacher.eu/question/other-ways-to-say-according-to/`).

[8] See our website at `http://www.hlt.utdallas.edu/~persingq/ICLE/` for our manually constructed list of hedge words.

path with an odd negation count, we turn on the Disagree Somewhat feature. Similarly, if we find a hedged stancetaking path with an even negation count, we turn on the Agree Somewhat feature.

3. (Dis)Agree Strongly. When we believe there is strong evidence that an instance should belong to one of the Strongly classes, we turn on the corresponding (Dis)Agree Strongly feature. In particular, if we find a relevant stancetaking path that appears to agree with the prompt part (as described in Section 4.1.2), but do not find any such path that appears to disagree with it, we turn on the Agree Strongly feature. Similarly, if we find a relevant stancetaking path that appears to disagree with the prompt part, but do not find a relevant stancetaking path that appears to agree with it, we turn on the Disagree Strongly feature.

5 Evaluation

5.1 Experimental Setup

Data partition. All our results are obtained via leave-one-prompt-out cross-validation experiments. So, in each fold experiment, we partition the instances from our 11 prompts into a training set (10 prompts) and a test set (1 prompt).

Evaluation metrics. We employ two metrics to evaluate our systems: (1) micro F-score, which treats each instance as having equal weight; and (2) macro F-score, which treats each class as having equal weight.[9] To gain insights into how different systems perform on different classes, we additionally report per-class F-scores.

Training. We train the baselines and our approach using two learning algorithms, MALLET's (McCallum, 2002) implementation of maximum entropy (MaxEnt) classification and our own implementation of the one nearest neighbor (1NN) algorithm using the cosine similarity metric. Note that these two learners have their own strengths and weaknesses: in comparison to 1NN, MaxEnt is better at exploiting high-dimensional features but less robust to skewed class distributions. For the baseline systems, we select the learner by performing cross validation on the training folds to maximize the average of micro and macro F-scores in each fold experiment.

When training our approach, we perform exhaustive feature selection to determine which subset of the four sets of features (i.e., n-gram, duplicated Faulkner, path-based, and knowledge-based features) should be used. Specifically, we select the feature groups and learner jointly by performing cross validation on the training folds, choosing the combination yielding the highest average of micro and macro F-scores in each fold experiment. To prevent any feature type from dominating the others, to each feature we apply a weight of one divided by the number of features having its type.

Testing. In case of a tie when applying 1NN, the tie is broken by selecting the class appearing higher in Table 2.

5.2 Results and Discussion

Results on fine-grained essay stance classification are shown in Table 3. The first four rows show our baselines' performances. Among the four baselines, Always Agree Strongly performs best w.r.t. micro F-score, obtaining a score of 55.6%, whereas Duplicated Faulkner performs best w.r.t. macro F-score, obtaining a score of 15.6%. Despite its poor performance, Duplicated Faulkner is a state-of-the-art approach on this task. Its poor performance can be attributed to three major factors. First, it was intended to identify only Agree and Disagree instances (note that Faulkner simply removed neutral instances from his experimental setup), which should not prevent them from performing well w.r.t. micro F-score. Second, it is far too permissive, generating features from a large majority of sentences while relevant sentences are far rarer. Third, while it does succeed at predicting Disagree Strongly far more frequently than either of the other baselines that excludes the Faulkner feature set, the problem's class skewness means that a learner is much more likely to be punished for predicting minority classes, which are more difficult to predict with high precision.

The fact that it makes an attempt to solve the problem rather than relying on class skewness for good performance makes Duplicated Faulkner a more interesting baseline than either N-Gram or Always Agree Strongly, even though both technically outperform it w.r.t. micro F-score. Similarly, the statistically significant improvements in micro and macro F-score our approach achieves over the best baselines are more impressive when taking the skewness problem into consideration.

The results of our approach, which has access

[9] Since stance classification is a multiclass, single-label task, micro F-score, precision, recall, and accuracy are all equivalent.

	System	Micro-F	Macro-F	A+	A−	Neu	D−	D+	Nev
1	Always Agree Strongly	55.6	11.9	71.4	.0	.0	.0	.0	.0
2	N-Gram	55.4	12.0	71.3	.0	.0	.0	.5	.0
3	Duplicated Faulkner	50.8	15.6	66.8	4.0	.0	.0	22.9	.0
4	N-Gram + Duplicated Faulkner	53.4	15.4	69.1	2.5	.0	.0	20.6	.0
5	Our approach	**60.6**	**20.1**	73.6	.0	.0	2.1	44.8	.0

Table 3: Cross-validation results for fine-grained essay stance classification, including per-class F-scores for Agree Strongly (A+), Agree Somewhat (A−), Neutral (Neu), Disagree Somewhat (D−), Disagree Strongly (D+), and Never Addressed (Nev).

to all four feature groups, are shown in row 5 of the table. It obtains micro and macro F-scores of 60.6% and 20.1%, which correspond to statistically significant relative error reductions over the best baselines of 11.3% and 5.3%, respectively.[10]

Recall that we turned on one of our knowledge-based features only when we believed there was strong evidence that an instance belonged to its associated class. To get an idea of how useful these features are, we calculate the precision, recall, and F-score that would be obtained for each class if we treated our knowledge-based features as heuristic classifiers. The respective precisions, recalls, and F-scores we obtained are: 0.66/0.28/0.40 (A+), 0.50/0.02/0.04 (A−), 0.00/0.00/0.00 (Neu), 0.50/0.01/0.02 (D−), and 0.63/0.31/0.42 (D+). Since the rule predictions are encoded as features for the learner, they may not necessarily be used by the learner even if the underlying rules are precise. For instance, despite the rule's high precision on the Agree Somewhat class, the learner did not make use of its predictions due to its low coverage.

5.3 Additional Experiments

Since all the systems we examined fared poorly on identifying Somewhat classes, one may wonder how these systems would perform if we considered a simplified version of the task where we merged each Somewhat class with the corresponding Strongly class. In particular, since Faulkner's approach was originally not designed to distinguish between Strongly and Somewhat classes, it may seem fairer to compare our approach against Duplicated Faulkner on the four-class essay stance classification task, where stance can take one of four values: Agree (created by merging Agree Strongly and Agree Somewhat), Disagree (created by merging Disagree Strongly and Disagree Somewhat), Neutral, and Never Addressed.

In the results for the different systems on this four-class stance classification task, shown in Table 4, we see that the same patterns we noticed in the six-class version of the task persist. The approaches' relative order w.r.t. micro and macro F-score remains the same, though they are adjusted upwards due to the problem's increased simplicity. Our approach's performance on Agree increases (compared to Agree Strongly) because Agree is a bigger class, making predictions of the class safer. Our approach's performance decreases on Disagree (compared to Disagree Strongly) since it is not good at predicting Disagree Somewhat instances which are part of the class.

5.4 Error Analysis

To gain additional insights into our approach, we analyze its six major sources of error below.

Stances not presented in a straightforward manner. As an example, consider "To my opinion this technological progress triggers off the imagination in a certain way." To identify this sentence as strongly disagreeing with the proposition "there is no longer a place for dreaming and imagination", we need to understand (1) the world knowledge that technological progress is occurring, (2) that "triggers off the imagination in a certain way" means that the technological progress coincides with imagination occurring, (3) that if imagination is occurring, there must be "a place for dreaming and imagination", and (4) that the prompt part is negated. In general, in order to construct reliable features to increase our coverage of essays that express their stance like this, we would need additional world knowledge and a deeper understanding of the text.

Rhetorical statements occasionally misidentified as stancetaking. For example, our method

[10] All significance tests are approximate randomization tests with $p < 0.01$. Boldfaced results are significant w.r.t. micro F-score for the Always Agree Strongly baseline, and macro F-score w.r.t. the Duplicated Faulkner baseline.

	System	Micro-F	Macro-F	A	Neu	D	Nev
1	Always Agree Strongly	64.8	19.7	78.7	.0	.0	.0
2	N-Gram	64.3	19.7	78.2	.0	.8	.0
3	Duplicated Faulkner	62.3	25.1	75.1	.0	25.2	.0
4	N-Gram + Duplicated Faulkner	62.6	23.7	75.8	.0	19.0	.0
5	Our approach	**67.6**	**29.1**	78.5	.0	38.0	.0

Table 4: Cross-validation results for four-class essay stance classification for Agree (A), Neutral (Neu), Disagree (D), and Never Addressed (Nev).

for identifying stancetaking paths misidentifies "I am going to discuss the topic that television is the opium of the masses in modern society" as stancetaking. To handle this, we need to incorporate more sophisticated methods for detecting rhetorical statements than those we are using (e.g., ignoring sentences ending in question marks).

Negation expressed without negation words. Our techniques for capturing negation are unable to detect when negation is expressed without the use of simple negation words. For example, "In this sense money is the root of life" should strongly disagree with "money is the root of all evil". The author replaced "life" with "evil", and detecting that this constitutes something like negation would require semantic knowledge about words that are somehow opposite of each other.

Insufficient feature/heuristic coverage of the Disagree Strongly class. Our stancetaking path-based features that we identified as intuitively having a connection to the Disagree Strongly class together cover only 51% of Disagree Strongly instances, meaning that it is in principle impossible for our system to identify the remaining 49%. However, our decision to incorporate only features that are expected to have fairly high precision for some class was intentional, as the lesson we learned from the Faulkner-based system is that it is difficult to learn a good classifier for stance classification using a large number of weakly or non-predictive features. To solve this problem, we would therefore need to exploit other aspects of strongly disagreeing essays that act as reliable predictors of the class.

Rarity of minority class instances. It is perhaps not surprising that our learning-based approach performs poorly on the minority classes. Even though the knowledge-based features were designed in part to improve the prediction of minority classes, our results suggest that the resulting features were not effectively exploited by the learners. To address this problem, one could employ a hybrid rule-based and learning-based approach where we use our machine-learned classifier to classify an instance only if it cannot be classified by any of these rules.

Lack of obvious similarity between instances of the same class. For example, if the most straightforward stancetaking sentence in an Agree Somewhat instance reads something like this, "In conclusion, I will not go to such extremes as to declare nihilistically that university does not prepare me for the real world in the least", (given the prompt part "Most university degrees do not prepare us for real life"), and we somehow managed to identify the instance's class as Agree Somewhat, what would the instance have in common with other Agree Somewhat instances? Given the numerous ways of expressing a stance, we believe a deeper understanding of essay text is required in order automatically detect how instances like this are similar to instances of the same class, and such similarities are required for learning in general.

6 Conclusion

We examined the new task of fine-grained essay stance classification, in which we determine stance for each prompt part and allow stance to take one of six values. We addressed this task by proposing two novel types of features, stancetaking path-based features and knowledge-based features. In an evaluation on 826 argumentative essays, our learning-based approach, which combines our novel features with n-gram features and Faulkner's features, significantly outperformed four baselines, including our re-implementation of Faulkner's system. Compared to the best baselines, our approach yielded relative error reductions of 11.3% and 5.3%, in micro and macro F-score, respectively. Nevertheless, accurately predicting the Somewhat, Neutral, and Never Addressed stances remains a challenging task. To stimulate further research on this task, we make all of our stance annotations publicly available.

Acknowledgments

We thank the three anonymous reviewers for their detailed comments. This work was supported in part by NSF Grants IIS-1219142 and IIS-1528037. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views or official policies, either expressed or implied, of NSF.

References

Amjad Abu-Jbara, Ben King, Mona Diab, and Dragomir Radev. 2013. Identifying opinion subgroups in arabic online discussions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 829–835, Sofia, Bulgaria, August. Association for Computational Linguistics.

Rakesh Agrawal, Sridhar Rajagopalan, Ramakrishnan Srikant, and Yirong Xu. 2003. Mining newsgroups using networks arising from social behavior. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 529–535, New York, NY, USA. ACM.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with E-rater v.2.0. *Journal of Technology, Learning, and Assessment*, 4(3).

Alexandra Balahur, Zornitsa Kozareva, and Andrés Montoyo. 2009. Determining the polarity and source of opinions expressed in political debates. In *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '09, pages 468–480, Berlin, Heidelberg. Springer-Verlag.

Mohit Bansal, Claire Cardie, and Lillian Lee. 2008. The power of negative thinking: Exploiting label disagreement in the min-cut classification framework. In *Coling 2008: Companion volume: Posters*, pages 15–18, Manchester, UK, August. Coling 2008 Organizing Committee.

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogs. In *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing*, ICSC '11, pages 162–168, Washington, DC, USA. IEEE Computer Society.

Filip Boltužić and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, Baltimore, Maryland, June. Association for Computational Linguistics.

Clinton Burfoot, Steven Bird, and Timothy Baldwin. 2011. Collective classification of congressional floor-debate transcripts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1506–1515, Portland, Oregon, USA, June. Association for Computational Linguistics.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. 2010. Probabilistic frame-semantic parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 948–956, Stroudsburg, PA, USA.

Adam Faulkner. 2014. Automated classification of stance in student essays: An approach using stance target information and the wikipedia link-based measure. In *Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference, FLAIRS 2014, Pensacola Beach, Florida, May 21-23, 2014.*

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English (Version 2)*. Presses universitaires de Louvain.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Ken Hyland. 2005. *Metadiscourse: Exploring interaction in writing*. Number London. Continuum.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. `http://mallet.cs.umass.edu`.

Akiko Murakami and Rudy Raymond. 2010. Support or oppose? classifying positions in online debates from reply activities and opinion expressions. In *Coling 2010: Posters*, pages 869–875, Beijing, China, August. Coling 2010 Organizing Committee.

Isaac Persing and Vincent Ng. 2015. 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)*, pages 543–552, Beijing, China, July. Association for Computational Linguistics.

Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From argumentation mining to stance classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 67–77, Denver, CO, June. Association for Computational Linguistics.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. 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)*, pages 116–125, Beijing, China, July. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar, October. Association for Computational Linguistics.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, Sydney, Australia, July. Association for Computational Linguistics.

Marilyn Walker, Pranav Anand, Rob Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 592–596, Montréal, Canada, June. Association for Computational Linguistics.

Yi-Chia Wang and Carolyn P. Rosé. 2010. Making conversational structure explicit: Identification of initiation-response pairs within online discussions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 673–676, Los Angeles, California, June. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2):0.

Ainur Yessenalina, Yisong Yue, and Claire Cardie. 2010. Multi-level structured models for document-level sentiment classification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1046–1056, Cambridge, MA, October. Association for Computational Linguistics.