# Measuring Online Debaters' Persuasive Skill from Text over Time

**Kelvin Luu**[1] **Chenhao Tan**[2] **Noah A. Smith**[1,3]
[1]Paul G. Allen School of Computer Science & Engineering, University of Washington
[2]Department of Computer Science, University of Colorado Boulder
[3]Allen Institute for Artificial Intelligence
{kellu,nasmith}@cs.washington.edu chenhao.tan@colorado.edu

## Abstract

Online debates allow people to express their persuasive abilities and provide exciting opportunities for understanding persuasion. Prior studies have focused on studying persuasion in debate content, but without accounting for each debater's history or exploring the progression of a debater's persuasive ability. We study *debater skill* by modeling how participants progress over time in a collection of debates from Debate.org. We build on a widely used model of skill in two-player games and augment it with linguistic features of a debater's content. We show that online debaters' skill levels do tend to improve over time. Incorporating linguistic profiles leads to more robust skill estimation than winning records alone. Notably, we find that an interaction feature combining uncertainty cues (hedging) with terms strongly associated with either side of a particular debate (fightin' words) is more predictive than either feature on its own, indicating the importance of fine-grained linguistic features.

## 1 Introduction

Persuasion is an important skill with prevalent use. Nearly every kind of social encounter, from formal political debates to casual conversations, can include attempts to convince others. What linguistic phenomena are associated with higher levels of persuasive skill? How do people develop this skill over time? Online debate communities offer an opportunity to investigate these questions. These communities feature users who participate in multiple debates over their period of engagement. Such debates involve two parties who willingly and formally present divergent opinions before an audience. Unlike other media of persuasion, such as letters to politicians, there is a clear signal, a win or loss, indicating whether or not a debater was successful against the adversary. This work aims to quantify the skill level of each debater in an online community and also investigates what factors contribute to expertise.

Although persuasion has generated interest in the natural language processing community, most researchers have not tried to quantify the persuasiveness of a particular speaker. Instead, they estimate how persuasive a *text* is, using linguistic features such as the author's choice of wording or how they interact with the audience (Tan et al., 2014, 2016; Althoff et al., 2014; Danescu-Niculescu-Mizil et al., 2012). Previous research has also established that debaters' content and interactions both contribute to the success of the persuader (Tan et al., 2016; Zhang et al., 2016; Wang et al., 2017). These works have found textual factors that contribute to a debater's success, but they do not emphasize the role of the individual debater.

There has been some recent work on studying individual debaters. Durmus and Cardie (2019) analyze users and find that a user's success and improvement depend on their social network features. We take another approach by estimating each user's skill level in each debate they participate in by considering their debate history. These estimates reveal features correlated with skill and the importance of particular debates over time. Our study is based on debates from an online debate forum, Debate.org, introduced by Durmus and Cardie (2018) and discussed in §2.

This Web site is composed of primarily text-based debates and attracts a large number of users to debate regularly.

Our model of skill builds on the Elo (1978) rating system, designed for rating players in two-player games (§3). Our preliminary analysis using Elo scores suggests that user skill is not static; debaters in the Debate.org forum tend



Transactions of the Association for Computational Linguistics, vol. 7, pp. 537–550, 2019. https://doi.org/10.1162/TACL_a_00281
Action Editor: Jordan Boyd-Graber. Submission batch: 11/2018; Revision batch: 4/2019; Published 9/2019.

to improve with practice. We extend Elo for the debate setting using linguistic features. We decompose our family of models into two design choices that align with the questions we hope to answer: 1) the features we use and 2) how we might choose to aggregate those features from past debates.

To validate our skill estimates, we introduce a **forecasting** task (§4). Previous work predicted the winner of a debate using the text of the current debate. In contrast, we aim to predict winners using our skill estimates *before* the debate (ignoring the current debate's content). This design ensures that we are modeling skill of the debater, as inferred from past performance, not the idiosyncrasies of a particular debate.

We also investigate the predictive power of our estimates through an analysis of the results (§6). We show that our full model outperforms the baseline Elo model, approaching the accuracy of an oracle that *does* use the text of the current debate. Our ablation studies reveal that the co-occurrence of phrases that indicate uncertainty or doubt (**hedges**) with words that are strongly associated with one debater or the other (**fightin' words**), is an effective predictor. Moreover, we find that not all past debates are equally useful for prediction: more recent debates are more indicative of the user's current level of expertise. This adds support to our conjecture that individual debaters tend to improve through the course of their time debating.

Finally, we track the linguistic tendencies of each debater over the course of their debating history. We show that several features such as the length of their turns and the co-occurrence of hedges and fightin' words increase over time for the best debaters, but stay static for those with less skill. These findings give further evidence that debaters improve over time.

## 2 Data

We use both debate and user skill data from the `Debate.org` data set introduced by Durmus and Cardie (2018). Any registered user on the Web site can initiate debates with others or vote on debates conducted by others.

### 2.1 Mechanism of `Debate.org`

Registered users can create a debate under a topic of their choosing. The person initiating the debate, called the **instigator**, fixes the debate's number of rounds (2–10) and chooses the category (e.g., politics, economics, or music) at the start of the debate. The instigator then presents an opening statement in the first round and waits for another user, the **contender**, to accept the debate and write another opening statement to complete the first round.[1] We define a debater's **role** as being either the instigator or contender.


Figure 1: An example of the `Debate.org` voting system.

To determine the debate winner, other `Debate.org` users vote after the debate ends. In this phase, voters mark who they thought performed better in each of seven categories (see Figure 1).[2] This phase can last between 3 days and 6 months, depending on the instigator's choice at the debate's creation. After this period, the debater who received the most points wins the debate. We record in our data set the textual information, participants, and voting records for each debate.

### 2.2 Definition of Winning

The `Debate.org` voting system lets us define a ''win'' in various ways (see Figure 1). In this study, we would like to model how convincing each debater is. One approach might be based on Oxford-style debates like the IQ2 data set from Zhang et al. (2016), in which scores are based on the number of audience members who changed their minds as a result of the debate. We found here that the majority of voters do not deviate from their stances before the debate, and most voters tend to vote for who they already agree

[1] Although many debaters use their first round to make an opening statement, some use it only to propose and accept debates. If the first turn in an $n$-round debate is under 250 words, we merge each debater's first two turns and treat the debate as an $(n-1)$-round debate.

[2] Another system of voting lets voters choose who they thought performed better over the entire debate. Although this appears in the data set, we do not use it in this paper.

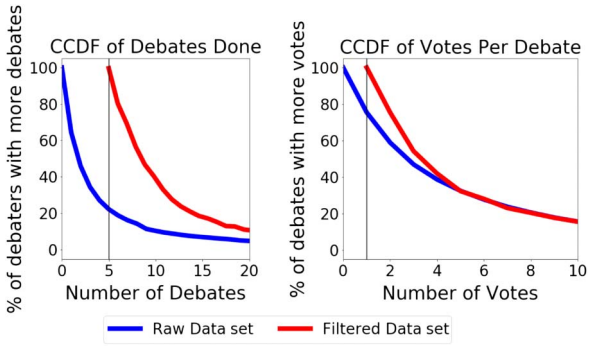


Figure 2: Complementary cumulative distribution functions (1 − CDF) for the total number of debates a user finished. The blue line tracks only debaters who engaged in and successfully concluded at least one debate, and the red line tracks users who have finished at least five debates in the filtered data set. The right plot similarly shows the complementary cumulative distribution for the number of votes given per debate for all debates and the filtered data set.

with. Therefore, we count the number of times each debater was rated as more convincing to a voter *despite* presenting a viewpoint that the voter disagreed with before the debate. The debater with the higher count of such votes is considered the ‘‘winner’’ in the remainder of this paper.

### 2.3 Data Set Statistics

From an unfiltered set of 77,595 debates, we remove all debates where a user forfeits, that lack a winner (§2.2), or that do not have a typical voting style with seven categories, leaving 29,209 completed debates.

We record the number of debates that each user completes (see Figure 2, left). We find that the quantity of debates per user follows a heavy-tailed distribution where most users do not participate in more than one debate. For the remainder of the work, we focus on the 1,284 (out of 42,424 total users) who have completed five or more of the debates described above. This leaves us with 4,486 debates where the participants have completed at least five debates (see Table 1).

We also record the number of votes that each debate attracted (see Figure 2, right). This number also follows a heavy-tailed distribution.

## 3 Expertise Estimation

In order to explore debater expertise and discover what contributes to a user's expertise over their time on `Debate.org`, we begin with a conventional approach to skill estimation, the Elo rating system, which serves as both a baseline and the basis for our final model. In our initial data analysis, we use Elo scores to define an **upset** as a debate where a weaker debater wins over a stronger one. By measuring the rate at which upsets occur over a debater's lifetime, we make an initial observation that debaters seem to improve with experience.

| | #Users | #Debates |
|---|---|---|
| Completed Debates | 42,424 | 77,595 |
| Completed & Convincing | 21,753 | 29,209 |
| Full Filtered | 1,284 | 4,486 |

Table 1: Description of the `debate.org` data set from Durmus and Cardie (2018) and the filtered data sets. **Full Filtered** is a subset of **Completed & Convincing** that requires that participants of each debate have engaged in five or more debates. We use **Full Filtered** for the remainder of our analysis.

### 3.1 Elo Model

Elo originated as a ranking system for chess players; it has been adapted to other domains, such as video games. It is one of the standard methods to rate players of a two-player, zero-sum game (Elo, 1978).[3] Elo assigns positive integer-valued scores, typically below 3,000, with higher values interpreted as ‘‘more skill.’’ The difference in the scores between two debaters under a logistic model is used as an estimate of the probability each debater will win. For example, consider a debate between $A$ and $B$. $A$ has an Elo rating of $R_A = 1900$, and $B$ has an Elo rating of $R_B = 2000$. Using the Elo rating system, $p_A$, the probability that $A$ wins is[4]

$$p_A = \frac{1}{1 + 10^{0.0025(R_B - R_A)}} = \frac{1}{1 + 10^{0.25}} \approx 0.36 \quad (1)$$

Ratings are updated after every debate, with the winner (equal to $A$ or $B$) gaining (and the loser losing) $\Delta = 32(1 - p_W)$ points. (32 is an arbitrary scalar; we follow non-master chess in selecting this value.)

Note that the magnitude of the change corresponds to how unlikely the outcome was.

[3]The Elo model is a special case of the Bradley-Terry model (Bradley and Terry, 1952).

[4]The base of the exponent, 10, and the multiplicative factor on the difference $R_A - R_B$, 0.0025, are typically used in chess.

Although the Elo ratings traditionally take only a win or loss as input, there have been adjustments to account for the magnitude of victory. One such method would be to use the score difference between the two players to adjust the Elo gain (Silver, 2015). If we let $S_A$ and $S_B$ be scores for $A$ and $B$, respectively, the modified gain $\Delta'$ is

$$\Delta' = \log(|S_A - S_B| + 1) \times \Delta \qquad (2)$$

Under this model, we represent a user's history and skill level as a single scalar, that is, their Elo rating. The Elo system ignores all other features, which include the style a debater uses in the debates and the content of their argument. We therefore view this model as a baseline and extend it.

### 3.2 Do Debaters Get Better Over Time?

Our initial data analysis uses the Elo model to investigate whether users improve over time at all in the first place. An increase in Elo score can be seen as the rating system merely becoming more accurate with another sample or an actual increase in the player's skill. We wish to show that there is no fixed rating for a user, but rather that the rating is a moving target. A counterhypothesis is then that the users' skill levels do not change despite their activity on `Debate.org`. If true, then we would expect their final Elo rating to be also indicative of their skill level at the beginning of their `Debate.org` activity.

To test this hypothesis, we first define an **upset** as a debate where $p_W < \tau$, that is, where the winner of the debate was estimated (under Elo) to win with low probability (set using threshold $\tau$). If users tend to have static skill levels, then participating in many debates does not affect debaters' skill, but simply provides more samples for measuring their skill. It follows that we would expect upsets to occur at the same rate early and late in each one's career given debaters' static skill. In this analysis, we calculate $p_A$ and $p_B$ for each debate using $A$'s and $B$'s *final* Elo ratings, which we take to be the most accurate estimate of their (presumed static) skill levels.[5]

To operationalize ''early'' and ''late,'' we divide a user's debates into quintiles by time, comparing the upset rate in different quintiles. In this analysis, we only consider users with at least ten debates ($N = 4,420$).[6]

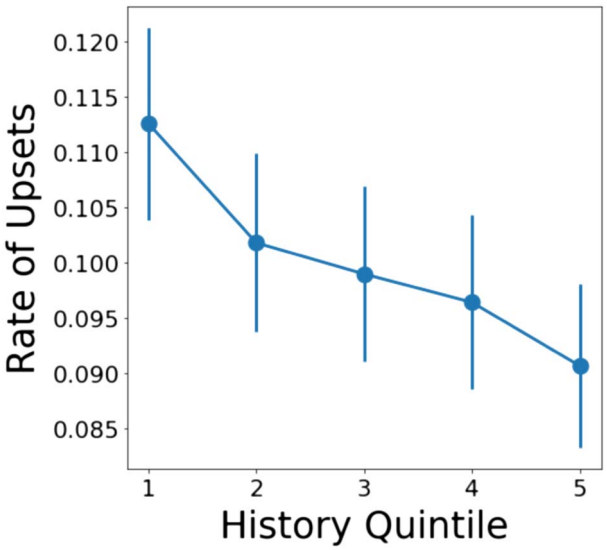


Figure 3: Upset rates (aggregated across users) across history quintiles, $\tau = 0.45$. The error bars represent the 95% confidence intervals.

We see a downward trend in the upset rate (Figure 3). In particular, the first and last quintiles show a statistically significant difference under a paired $t$-test ($p < .001$), meaning that a user's final Elo score is not a good measure of skill at earlier times. We take this finding as suggestive that users of `Debate.org` adapt as they participate in more debates.

## 4 Predicting Expertise using Earlier Debates

Our aim is to estimate a debater $A$'s persuasive ability after observing a series of debates they participated in (denoted $d_1^A, \ldots, d_{t-1}^A$ if we are estimating ability just before the $t$th debate). We wish to take into account the content of those debates (not merely their outcomes, as in Elo), so as to understand what factor reveals a debater's skill levels. Drawing inspiration from Elo's interpretation as a score that can be used to predict each debater's probability of winning (Eq. 1), we formulate a prediction task: estimate $p_A$ for $A$'s $t$th debate, given $d_1^A, \ldots, d_{t-1}^A$ and also the opponent's debate history (which might be of a different length).

[5] In a forecasting analysis like the one in §4, this would be inappropriate, as it uses ''future'' information to define upsets. This notion of upset is not used in the forecasting task.

[6] In addition to requiring 10 debates instead of 5 per user, as in the rest of our analysis, we also do not subject the opponents to the same requirements.

By observing debate outcomes alongside the two participants' histories, we can estimate the parameters of such a probability model. Elo provides a baseline; rather than opaque scores associated with individual users at different times, we seek to *explain* the probability of winning through linguistic features of past debate content.

Unlike previous work (Zhang et al., 2016; Potash and Rumshisky, 2017; Tan et al., 2018), we do not use the content of the *current* debate ($d_t^A$) to predict its outcome; rather, we forecast the outcome of the debate as if the debate has not yet occurred. To our knowledge, this is the first work to derive scores for current skill levels based on observing participants' behavior over time. In the remainder of this section, we discuss features of past debates and ways of aggregating them.

### 4.1 Incorporating a Linguistic Profile into Elo

Elo scores are based entirely on wins and losses; they ignore debate content. We seek to incorporate content into expertise estimation by using linguistic features. If we modify the exponential base in Equation 1 from 10 to $e$, we can view Elo probabilities (e.g., $p_A$) as the output of a logistic regression model with one feature (the score difference, $R_A - R_B$) whose weight is 0.0025; that is, $p_A = \sigma(0.0025 \cdot (R_A - R_B))$. It is straightforward to incorporate more features, letting

$$p_A = \sigma\left(\mathbf{w} \cdot (\boldsymbol{R}_A - \boldsymbol{R}_B)\right) \tag{3}$$

where $\mathbf{w}$ is a vector of weights and $\boldsymbol{R}_U$ is user $U$'s ''profile,'' a vector of features derived from past debates. In this work, the linguistic profiles are designed based on extant theory about the linguistic markers of persuasion, and the vectors are *weighted averages* of features derived from earlier debates.

### 4.2 Features

We select features discussed in prior work as the basis for our linguistic profiles (Tan et al., 2016, 2018; Zhang et al., 2016). We extract these measurements from each of the user's debates. For a given debate and user, we calculate these values over the rounds written by the user. For example, if we were interested in a debate by Alex as the instigator, we would only calculate features from the instigator rounds of that debate (since the contender rounds of that debate were written by his opponent). Table 2 shows the full list of features.

**Hedging with fightin' words.** We introduce one novel feature for our work: the hedging with fightin' words. ''Fightin' words'' refer to words found using a method, introduced by Monroe et al. (2008), which seeks to identify words (or phrases) most strongly associated with one side or another in a debate or other partisan discourse.[7] We are interested in situations where debaters evoke fightin' words (their own, or their opponents') with an element of uncertainty or doubt. We use each debater's top 20 fightin' words (unigrams or bigrams) as features, following Zhang et al., 2016, who found this feature useful in predicting winners of Oxford-style televised debates. We also count cooccurences of fightin' words with hedge phrases like ''it could be possible that'' or ''it seems that.'' An example of this conjoined feature is found in the utterance ''Could you give evidence that **supports** the idea that married couples are **more likely** to be committed to [other tasks]?'', where hedge phrases are emboldened and brackets denote fightin' words (which are selected separately within each debate). We use a list of hedging cues curated by Tan et al. (2016) and derived from Hyland (1996) and Hanauer et al. (2012). The conjoined feature is the count of the user's sentences in a debate where a fightin' word cooccurs with a hedge phrase in a sentence.

### 4.3 Aggregating Earlier Debates

Because we consider the full history of a debater when estimating their skill level, we opt to aggregate the textual features over each debate. We do so by taking a weighted sum of the feature vectors of the previous debates. We consider four weighting schemes, none of which have free parameters, to preserve interpretability. Let $f$ be any one of the features in the linguistic profile, a function from a single debate to a scalar.

1. **Exponential growth:** The most recent debates are most indicative of skill, $\sum_{i=1}^{t-1} \frac{f(d_i^A)}{2^{t-i}}$. We take this to be the most intuitive choice, experimentally comparing against the alternatives below to confirm this intuition.

[7] The method estimates log-odds of words given a side, with Dirichlet smoothing, and returns the words with the highest log-odds for each side.

| Feature | Description | |
|---|---|---|
| Elo Score | Traditional Elo score calculated and updated. Updated traditionally, not averaged as in §4.3. | n/a |
| Length | Number of words this user uttered in the debate. | ↑↑↑ |
| Part of speech | Count of each noun, verb, adjective, preposition, adverb, or pronoun from the participant in the entire debate. | Nouns (↑↑↑) Adjectives (↑↑↑) |
| Flesch reading ease | Measure of readability given the number of sentences in a document and the number of words in each sentence (Kincaid et al., 1975). | ↑↑ |
| Emotional words | Cues that indicate a positive or negative emotion (Tausczik and Pennebaker, 2010). | Pos (↑↑↑) Neg (↑↑↑) |
| Links | Links to external websites outside of `debate.org`. This feature operationalizes the number of sources a debater used. | |
| Questions | The number of questions the user asked in the debate. | ↓↓↓ |
| Quotations | The number of quotations the user included in the debate. | |
| Hedging | The number of phrases that soften a statement by adding uncertainty (Hyland, 1996; Hanauer et al., 2012). | |
| Fightin' words | The number of instances of words most strongly associated with either debater (Monroe et al., 2008). | ↑↑↑ |
| H∧FW | The number of cooccurences of hedging and fightin' words, described in §4.2. | ↑↑ |

Table 2: Debate-level features used in estimating skill levels. Aside from Elo, the features are a part of the user's linguistic profile. The third column represents statistical significance levels in comparing winners and losers' features (independently) with Bonferroni correction: ↑ is $p < 0.05$, ↑↑ is $p < 0.01$, ↑↑↑ is $p < 0.001$.

2. **Simple average:** Each earlier debate's feature vector is weighted equally, $\frac{1}{t-1}\sum_{i=1}^{t-1} f(d_i^A)$.

3. **Exponential decay:** The first debates are most indicative of skill, $\sum_{i=1}^{t-1} \frac{f(d_i^A)}{2^i}$.

4. **Last only:** Only the single most recent debate matters, $f(d_{t-1}^A)$ (an extreme version of ''exponential growth'').

In each variation of our method, some or all of the linguistic profile features are aggregated (using one of the four weighted averages), then applied to predict debate outcomes through logistic regression. We note that if our sole aim were to maximize predictive accuracy, we might explore much richer linguistic profiles, perhaps learning word embeddings for the task and combining them using neural networks and enabling interactions between a debate's two participants' profiles.[8] In this study, we seek to estimate skill but also to understand it, so our focus remains on linear models.

## 5 Experimental Setup

We validate our skill models with a binary classification task, specifically forecasting which of two debaters will win a debate (without looking at the content of the debate). Here we remove any debates where someone forfeits or there is no winner. We then considered debates where each debater has completed at least five of the remaining debates. As discussed in §2.2, the winner is taken to be the debater receiving the most ''more convincing argument'' votes from observers who did not initially agree with them. We create four training/evaluation splits of the data, using debates from 2013, 2014, 2015, and 2016 as evaluation sets (i.e., development and test) and debates prior to the evaluation year for training. Figure 4 shows the number of debates in each split. We note that our training sets are cumulative. For example, if we were to test on 2015 data, we would use the 2012, 2013, and 2014 training as training data.

Because we do not test on 2012 data or train on 2016 data due to the low number of debates before 2012 and after 2016, we treat the whole of 2012 as training data and 2016 as development and test. We report the accuracy for each run.

We compare several predictors:[9]

- **Full model**: Our model with all features, as described in §4, and (except where otherwise

[8] Indeed, in preliminary experiments we *did* explore using a recurrent neural network instead of a fixed weighted average, but it did not show any benefit, perhaps owing to the relatively small size of our data set.

[9] We use $\ell_2$ regularization in our models with the linguistic profile features.

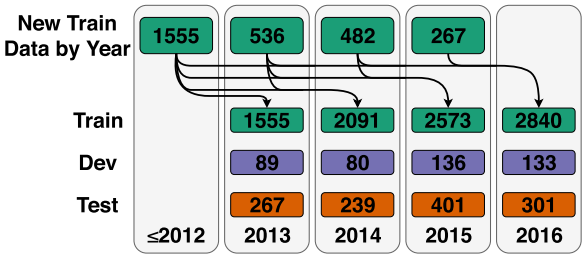


Figure 4: We split data into training/development/test based on year. This chart shows the number of debates in each subset of the data. We note that, for training, we use all the training debates from previous years (e.g., if we were to test on 2015, we would train using the training splits from 2012, 2013, and 2014) Each instance in this figure corresponds to a debater.

stated) the exponential growth weighting. This model combines linguistic profiles from earlier debates with a conventional Elo score.

- **Full model with point difference**: Our full model as described above, but we scale the Elo gain by the point difference as described in Equation 2.

- **Linguistic profile only**: Our model with exponential growth weighting (except where otherwise stated), but ablating the Elo feature. This model is most similar to those found in prior literature (Zhang et al., 2016; Tan et al., 2018; Wang et al., 2017).

- **Elo**: The prediction is based solely on the Elo score calculated just before the debate. This is equivalent to ablating the linguistic profiles from our model.

- **Final Elo oracle**: The prediction is based solely on the two debaters' *final* Elo scores (i.e., using all debates from the past, present, and future).

- **Current debate text oracle**: A model that uses the linguistic profile derived just from the current debate. Although this model is most similar to previous work, it is not a fair estimate of skill (because it ignores past performance). We therefore view it as another oracle.

- **Majority choice**: A baseline that always predicts that the contender will win.[10]

[10]In this data set, contenders win nearly 59% of the time, a fact frequently discussed in the Debate.org community; see http://ddo.wikia.com/wiki/Contender_Advantage. The contender advantage is sometimes attributed to having the "final say," or to the fact that contenders choose the instigators they wish to debate.

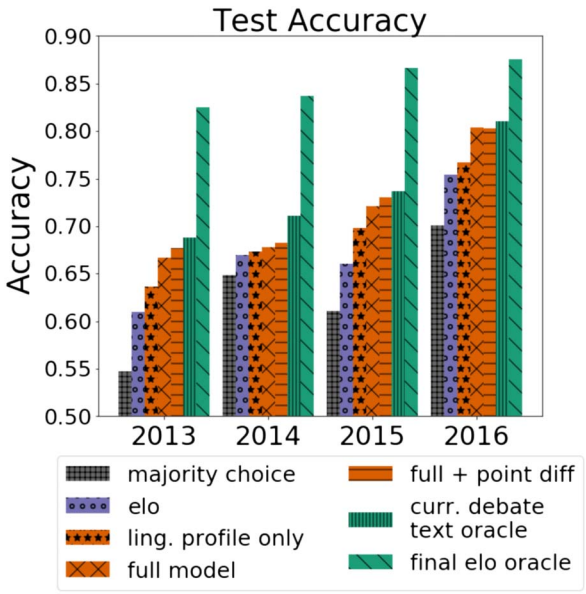


Figure 5: Our results for the prediction task. Our full model outperforms the Elo baseline and approaches the current debate text oracle.

## 6 Results

In this section, we first show that the expertise of a debater can be better estimated with the linguistic profile, and then analyze the contribution of different components. We further examine the robustness of our results by controlling for additional variables.

### 6.1 Prediction Performance

We first present our results with what we consider our best model, that is, our full model (with point differences), which consists of all features and uses the exponentially growing weight.

**Importance of linguistic features.** We see from Figure 5 that the full model outperforms the standard Elo baseline. The gap between the two models suggests that the addition of the linguistic profile contributes to the performance of the model and therefore plays a useful role in skill estimation. Moreover, the linguistic profile only model shows that the linguistic profile features are not only useful, but have at least as strong predictive power as Elo alone. By only using the linguistic features aggregated over the course of a debater's history, *without knowing winning records*, we can forecast at least as well as the Elo baseline.

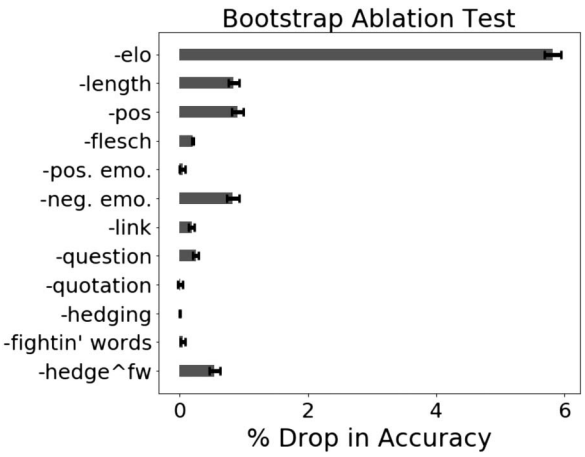


Figure 6: Our bootstrap on the feature ablations. We record the average drop in performance across 100 iterations and tested on the 2016 test set. Higher means a larger drop in performance.

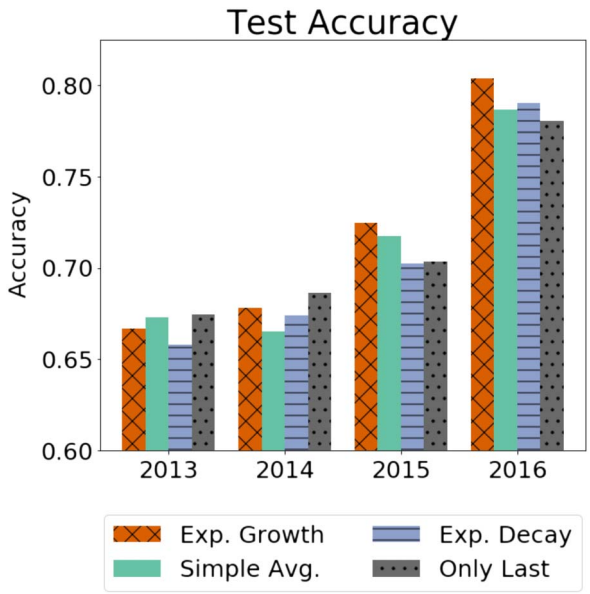


Figure 7: Comparison in performance for the four ways we aggregate features over time.

**Importance of multiple debates.** We also note that our full model only performs slightly worse than the current debate text oracle despite the current debate text model directly observing the content of the debate. This result implies that using information from only previous debates has at least similar predictive strength to information from the debate at hand. Moreover, the large gap between the final Elo and current debate text oracles implies that a user's skill is evidenced by more than the content of a single debate. These results further demonstrate the importance of accounting for debaters' prior history.

**Magnitude of victory might matter.** Our full model with the point difference scaled Elo gain does roughly as well or slightly better than our normal full model. As the focus of this paper is on incorporating linguistic profiles, we use the **full model without the point difference scaling** for analyses in the rest of the paper.

### 6.2 Feature Ablations

We inspect the contribution of each feature by removing each one from the model. Then, for each feature, we perform a bootstrap test over the last year of data (trained on 2012–2015 data; tested on 2016). At each iteration, we sample 1,000 training examples to train on, but fix the test set across iterations. We then train our full model alongside several other models, each with a feature ablated, on the sample. We track the drop in performance between our full model and each of our other models. We record the average performance over 100 iterations for comparison.

From Figure 6, we find that removing Elo results in the most severe drop in performance (5.8%). Ablating part-of-speech, negative emotion, and length from our model had a moderate effect on performance. Surprisingly, we find that, although the H∧FW feature is the overlap between hedge cues and fightin' words, the latter two features contribute very little to the performance of our model compared with H∧FW.

### 6.3 Combining Prior Debate Features

As described in §4.3, we explore several ways of combining features over the past debates. By inspecting how these different aggregation functions might differ in performance, we hope to find out whether or not some debates are more important than others, if recency matters at all, and if some history is important at all. We do so by 1) giving the last debates more weight (**exponential growth**), 2) giving all weights equal weight (**simple average**), 3) giving the first debates more weight (**exponential decay**), and 4) giving all the weight to the last debate (**last debate only**). The rest of this section discusses these results, shown in Figure 7.

**More debates help.** When using only the last debate's features and ignoring all previous debates, our performance is initially very good in the years 2013 and 2014 (when our training sets are

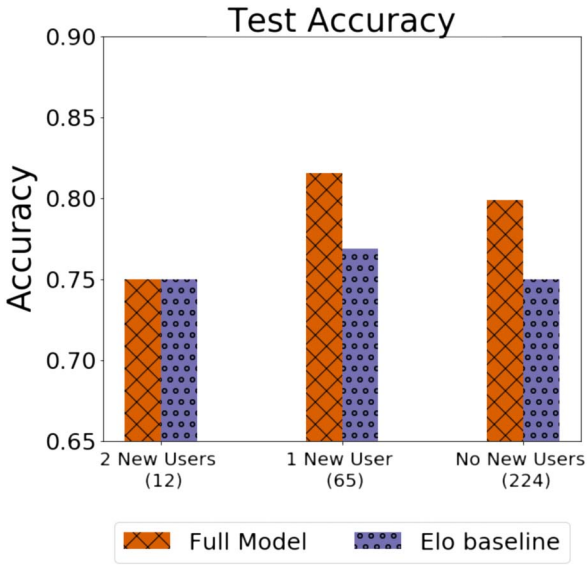


Figure 8: Our results separated by the number of times we have seen each participant before in training for the 2016 test set. The parenthetical numbers show how many debates are in the subset.

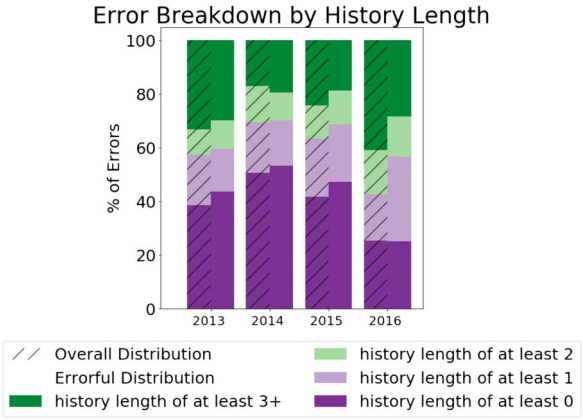


Figure 9: Errors of the full model broken down by the history sizes of the users. The overall distribution is to the left and the error distribution is to the right.

smaller and histories shorter). However, as debate histories become richer in the later years, last-only's performance drops in comparison to that of the growing weight aggregation. These results match our intuition: with more experienced users, considering more debates give us a better gauge of how a debater performs and a debater's more recent debates give a better snapshot of the user's current skill level.

**Later debates matter more.** As hinted in our last-only results, giving the last debates more weight does best in all four years. In light of its performance compared with the simple mean and decaying weight settings, our results imply that not all debates contribute equally to a debater's skill under our model. Indeed, our results show that the most recent debates are also the most important for estimating a debater's skill rating.

### 6.4 Impact of the Length of History

We finally examine how often our models have seen a user's history impacts performance. Compared with the **final Elo** and **current debate text** oracle models, both the **full** and the **Elo** models suffer when both debaters have no prior history—these examples devolve into having to guess the majority class.

Therefore, we further analyze these models by inspecting the performance on subsets of the test set pertaining to instances where 1) both users are new, 2) there is only one new user, and 3) both debaters have been seen before. We expect that we will have better performance as our debates have more complete debate histories. We complete the analysis by looking at the 2016 test.

Figure 8 shows that both models vastly improve when removing debates where users have no history and only improve as we require users to have longer histories. These results are consistent with our intuition that longer debate histories allow our models to better infer a debater's skill. Moreover, our model generally outperforms the Elo baseline across the different history lengths, implying that our model's performance is consistent. We do note, however, that the gap (between the full model and Elo baseline) is consistent when there is only one new user and no new users. This could imply that our model, in addition to using longer histories more effectively, is more capable of estimating a debater's skill against an unknown user compared with the normal Elo baseline. The slightly higher performance in the one-new-user case, compared with no-new-users, is likely due to the small size of each subset.

We extend our analysis of the effect of history lengths on skill estimation by inspecting the errors that our model makes. We are particularly interested in characterizing the mistakes that our model makes by the participant with the *shorter* debate history. For each debate that the full model predicts incorrectly, we keep count of history length of the user with the lower history count (e.g., if a debate has a user with a debate history of length 5 against a user with a debate history

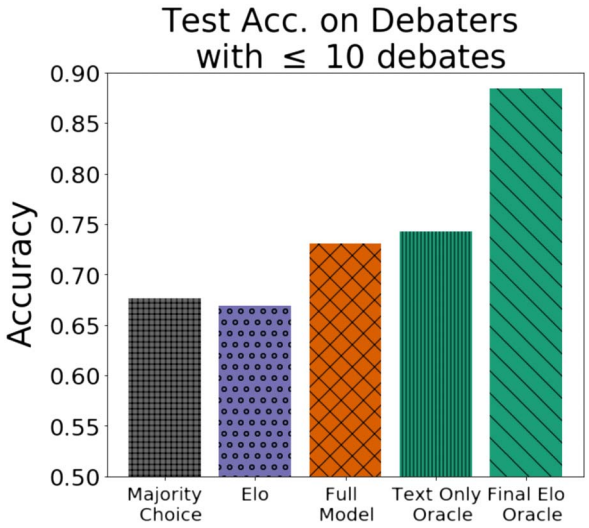


Figure 10: Our results for the prediction task on a restricted subset of the test set comprising only debates that are within both participants' first 10 debates. Although our model does suffer in accuracy under this setting, it still outperforms both baselines and approaches the text-only oracle.

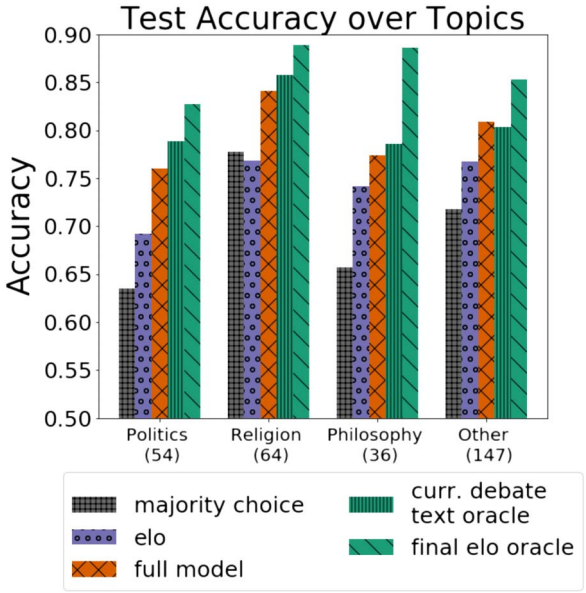


Figure 11: Our results for the prediction task broken down by topic of debate. Our full model maintains a consistent performance across all categories.

of length 1, we would treat the debate having a history size of 1). We find in Figure 9 that a large percentage of the errors (51 % across all debates) come from debates where one user had a history length of 0 at test time. Moreover, the debates with at least one user with no history tend to contribute proportionally more to the error rate relative to the overall distribution. Conversely, the debates with users who have deeper histories of three debates or more tend to contribute less to the error rate.

### 6.5 Controlling for Other Variables

We also inspect how well our model generalizes by measuring our model's accuracy on subsets of the test set. In particular, we hope to see if high activity users or the topic of a debate affect our model's performance.

**Controlling for prolific users.** As seen in Figure 2, debater activity follows a heavy-tailed distribution where a minority of debaters perform the majority of debates. In our data set, 18.9% of debaters who have engaged in more than 10 debates participate in 66.9% of the debates. Thus, in order to validate that our model does not simply recognize the most active users, we test on a subset of the 2016 data where we cap the number of debates a user can participate in to 10 (142 debates).

From Figure 10, we see that the order is roughly the same even when we restrict the test set to require that both users have participated in fewer than 10 debates at the time of the debate. This implies that our model generalizes relatively well to most debaters regardless of their experience. We do note, however, that most of the models incur a large decrease in performance whereas the final Elo oracle maintains roughly the same accuracy.

**Our observations are invariant to debate topic.** We also break down our test set by debate topic. We use the three most popular topics (politics, religion, and philosophy) and aggregate the rest of the categories as 'other'.

We see from Figure 11 that our model consistently outperforms both baselines and typically below the oracle models with the exception of 'other'. These results are consistent with those found in Figure 5 and show that our model performs consistently across different debate topics.

## 7 Language Change over Time: The Best Improve and the Worst Stagnate

Our findings from §3.2 and §6.3 imply that users tend to experience some change over time and that these changes are helpful for forecasting who in a debate is more likely to win. In this section, we explore whether debaters' linguistic tendencies

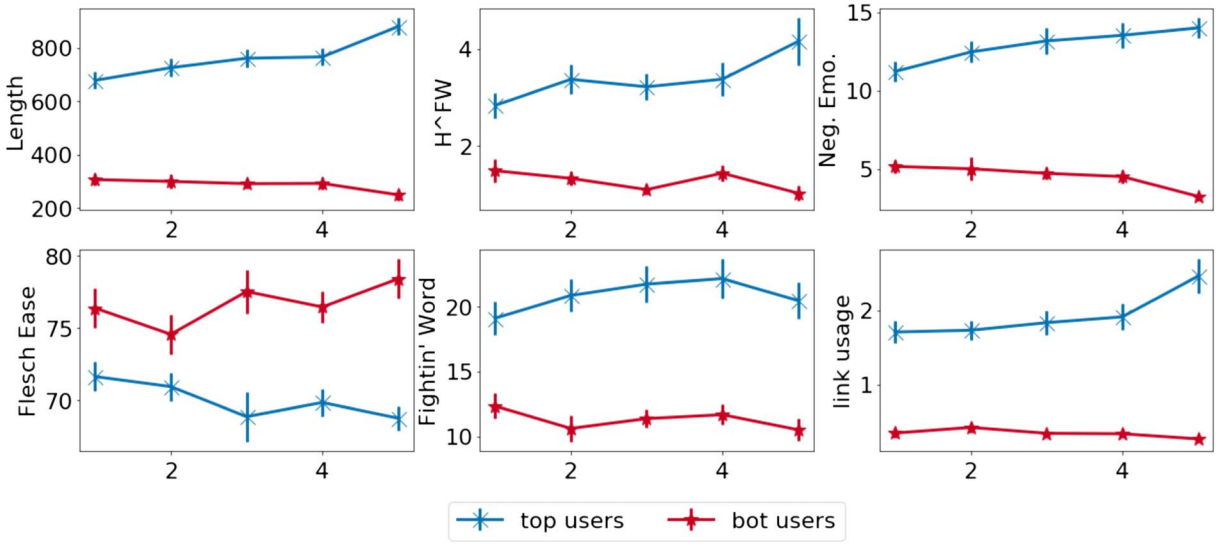


Figure 12: Feature measurements averaged across each history quintile. We see that there is a general trend for those who eventually become the best debaters to improve on these measures while the bottom users stagnate. The error bars represent the 95% confidence intervals.

change over time by tracking their use of features over the course of their debate history.

We examine how language use changes over time for the best and worst users. To do so, we divide each user's debate history into quintiles as in §3.2. For each quintile, we average the same features we use in the linguistic profile that we use in our model (Table 2). We take the top 100 and bottom 100 users ranked by our model to see how the trajectories change over time.

Figure 12 shows that best and worst users have different linguistic preferences even from the beginning of their debating activities. The best debaters have a higher feature count in every case except for the Flesch reading ease score. However, the best users do seem to improve over time in length, H∧FW, use of emotional cues, and link use, which mostly correlates with our feature ablation in Figure 6 ($p < 0.05$ after Bonferroni correction). In contrast, the worst users do not seem to experience any significant change over time except for the length of their rounds and negative emotional cue use. In those cases, the worst users seem to worsen over time.

## 8 Related Work

In addition to the most relevant studies mentioned so far, our work is related to three broad areas: skill estimation, argumentation mining, and studies of online debates.

**Skill estimation.** Ranking player strength has been studied extensively for sports and for online matchmaking in games such as *Halo* (Herbrich et al., 2007). The Bradley-Terry models (of which Elo is an example) serve as a basis for much of the research in learning from pairwise comparisons (Bradley and Terry, 1952; Elo, 1978). Another rating system used for online matchmaking is Microsoft's Trueskill rating system (Herbrich et al., 2007), which assumes performance is normally distributed. Neural networks have recently been explored (Chen and Joachims, 2016; Menke and Martinez, 2008; Delalleau et al., 2012), incorporating player or other contextual game features from previous games at the cost of interpretability of those features.

**Argumentation and persuasion.** Past studies have noted the persuasiveness of stylistic effects such as phrasing or linguistic accommodation. For example, Danescu-Niculescu-Mizil et al. (2012) showed that, in a pool of people vying to become an administrator of a Web site, those who were promoted tended to coordinate more than those who were not. Similarly, other work defines and discusses power relations over discussion threads such as emails (Prabhakaran and Rambow, 2014, 2013). Additionally, Tan et al. (2018) explored how debate quotes are selected by news media. They found that linguistic and interactive factors of an utterance are predictive of whether or not it would be quoted. Prabhakaran et al. (2014)

also studied political debates and found that a debater's tendency to switch topics correlates with their public perception. Argumentation has also been studied extensively in student persuasive essays and web discourse (Persing and Ng, 2015; Ong et al., 2014; Song et al., 2014; Stab and Gurevych, 2014; Habernal and Gurevych, 2017; Lippi and Torroni, 2016). Most relevant to our work on how users improve over time, Zhang et al. (2017) study how one document may improve over time through annotated revisions. Where our work examines users' linguistic change across multiple debates, they focus on how a user improves a single document over multiple revisions.

**Online debates.** There has also been recent work in characterizing specific arguments in online settings in contrast to our focus on the debaters themselves. For example, Somasundaran and Wiebe (2009), Walker et al. (2012), Qiu et al. (2015), and Sridhar et al. (2015) built systems for identifying the *stances* users take in online debate forums. Lukin et al. (2017) studied how persuasiveness of arguments depends on personality factors of the audience.

Other researchers have focused on annotation tasks. For example, Park and Cardie (2014) annotated online user comments to identify and classify different propositions. Hidey et al. (2017) annotated comments from the *changemyview* subreddit, a community where participants ask the community to change a view they hold. Likewise, Anand et al. (2011) annotated online blogs with a classification of persuasive tactics. Inspired by Aristotle's three modes of persuasion (ethos, pathos, and logos) their work annotates claims and premises within the comments. Habernal and Gurevych (2016) used crowdsourcing to study what makes an argument more convincing. They paired two similar arguments and asked annotators which one was more convincing. This framework allowed them to study the flaws in the less convincing arguments. The annotations they produced offer a rich understanding of arguments which, though costly, can be useful as future work.

## 9 Conclusion

In this work, we introduced a method that uses a linguistic profile derived from a debater's history of past debates to model their skill level as it changes over time. Using data from `Debate.org`, we formulate a forecasting task around predicting which of two debaters will be most convincing to observers predisposed to be unconvinced. We find that linguistic profiles on their own are similarly predictive to the classic Elo method, which does not parameterize skill according to attributes of a participant or their behavior, but only models wins and losses. Moreover, we show that our findings are robust to topic of debate and frequency of user activity. Further, a model combining linguistic profiles with Elo achieves predictive accuracy nearly on par with an oracle based on the content of the debate itself. A particular feature combining hedging with fightin' words is notably important in our model, and consistent with evidence that debaters improve with practice, more recent debates appear to provide better estimates of skill than earlier ones. To verify our hypothesis that users improve, we explicitly track the feature use of debaters over the course of their debating activity to show that the best users improve while the bottom users stagnate. Our approach sets the stage for future explorations of the role of history profile, discourse, more fine-grained sentiment features, or notions of topic in persuasion.

## Acknowledgments

We thank Esin Durmus for her help with the DDO corpus. We also thank the anonymous reviewers and the action editor for their helpful comments and suggestions that helped improve the paper. This research was supported in part by a University of Washington Innovation Award to the third author.

## References

Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2014. How to ask for a favor: A case study on the success of altruistic requests. In *Proceedings of ICWSM*.

Pranav Anand, Joseph King, Jordan Boyd-Graber, Earl Wagner, Craig Martell, Doug Oard, and Philip Resnik. 2011. Believe me? We can do this! Annotating persuasive acts in blog text. In *Proceedings of the Workshops at AAAI*.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Shuo Chen and Thorsten Joachims. 2016. Predicting matchups and preferences in context. In *Proceedings of SIGKDD*.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of WWW*.

Olivier Delalleau, Emile Contal, Eric Thibodeau-Laufer, Raul Chandias Ferrari, Yoshua Bengio, and Frank Zhang. 2012. Beyond skill rating: Advanced matchmaking in ghost recon online. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(3):167–177.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. In *Proceedings of NAACL-HLT*.

Esin Durmus and Claire Cardie. 2019. Modeling the factors of user success in online debate. In *Proceedings of WWW*.

Arpad E. Elo. 1978. *The Rating of Chessplayers, Past and Present*. Arco Publishers.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in Web argumentation. In *Proceedings of EMNLP*.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated Web discourse. *Computational Linguistics*, 43(1):125–179.

David A. Hanauer, Yang Liu, Qiaozhu Mei, Frank J. Manion, Ulysses J. Balis, and Kai Zheng. 2012. Hedging their mets: The use of uncertainty terms in clinical documents and its potential implications when sharing the documents with patients. In *Proceedings of AMIA*.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. TrueskillTM: A Bayesian skill rating system. In *Proceedings of NIPS*.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the Workshop on Argument Mining*.

Ken Hyland. 1996. Writing without conviction? Hedging in science research articles. *Applied Linguistics*, 17(4):433–454.

J Peter Kincaid, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. *CNTECHTRA Research Branch Report 8-75*.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):10.

Stephanie M. Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of EACL*.

Joshua E. Menke and Tony R. Martinez. 2008. A Bradley–Terry artificial neural network model for individual ratings in group competitions. *Neural Computing and Applications*, 17(2):175–186.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Nathan Ong, Diane Litman, and Alexandra Brusilovsky. 2014. Ontology-based argument mining and automatic essay scoring. In *Proceedings of the Workshop on Argumentation Mining*.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the Workshop on Argumentation Mining*.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of ACL*.

Peter Potash and Anna Rumshisky. 2017. Towards debate automation: A recurrent model for predicting debate winners. In *Proceedings of EMNLP*.

Vinodkumar Prabhakaran, Ashima Arora, and Owen Rambow. 2014. Staying on topic: An

indicator of power in political debates. In *Proceedings of EMNLP*.

Vinodkumar Prabhakaran and Owen Rambow. 2013. Written dialog and social power: Manifestations of different types of power in dialog behavior. In *Proceedings of IJCNLP*.

Vinodkumar Prabhakaran and Owen Rambow. 2014. Predicting power relations between participants in written dialog from a single thread. In *Proceedings of ACL*.

Minghui Qiu, Yanchuan Sim, Noah A. Smith, and Jing Jiang. 2015. Modeling user arguments, interactions, and attributes for stance prediction in online debate forums. In *Proceedings of SDM*.

Nate Silver. 2015. How our NFL predictions work. `https://fivethirtyeight.com/methodology/how-our-nfl-predictions-work/`. Accessed: 2019-02-30.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of ACL-IJCNLP*.

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the Workshop on Argumentation Mining*.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *Proceedings of ACL*.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of EMNLP*.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic-and author-controlled. In *Proceedings of ACL*.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of WWW*.

Chenhao Tan, Hao Peng, and Noah A. Smith. 2018. You are no Jack Kennedy: On media selection of highlights from presidential debates. In *Proceedings of WWW*.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Marilyn A. Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of NAACL-HLT*.

Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. 2017. Winning on the merits: The joint effects of content and style on debate outcomes. *Transactions of the Association for Computational Linguistics*, 5:219–232.

Fan Zhang, Homa B. Hashemi, Rebecca Hwa, and Diane Litman. 2017. A corpus of annotated revisions for studying argumentative writing. In *Proceedings of ACL*.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in Oxford-style debates. In *Proceedings of NAACL-HLT*.