In Opinion Holders' Shoes: Modeling Cumulative Influence for View Change in Online Argumentation

Zhen Guo
North Carolina State University
zguo8@ncsu.edu

Zhe Zhang
IBM Corporation
zhangzhe@us.ibm.com

Munindar P. Singh
North Carolina State University
singh@ncsu.edu

ABSTRACT

Understanding how people change their views during multiparty argumentative discussions is important in applications that involve human communication, e.g., in social media and education. Existing research focuses on lexical features of individual comments, dynamics of discussions, or the personalities of participants but deemphasizes the cumulative influence of the interplay of comments by different participants on a participant's mindset. We address the task of predicting the points where a user's view changes given an entire discussion, thereby tackling the confusion due to multiple plausible alternatives when considering the entirety of a discussion.

We make the following contributions. (1) Through a human study, we show that modeling a user's perception of comments is crucial in predicting persuasiveness. (2) We present a sequential model for cumulative influence that captures the interplay between comments as both local and nonlocal dependencies, and demonstrate its capability of selecting the most effective information for changing views. (3) We identify contextual and interactive features and propose sequence structures to incorporate these features. Our empirical evaluation using a Reddit Change My View dataset shows that contextual and interactive features are valuable in predicting view changes, and a sequential model notably outperforms the nonsequential baseline models.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; • **Theory of computation** → *Structured prediction*; • **Human-centered computing** → Social network analysis.

KEYWORDS

Online discussion modeling, Persuasion, Social media

ACM Reference Format:
Zhen Guo, Zhe Zhang, and Munindar P. Singh. 2020. In Opinion Holders' Shoes: Modeling Cumulative Influence for View Change in Online Argumentation. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3366423.3380302

1 INTRODUCTION

Argumentative discussions are common in daily life and on social media. In settings of interest here, the participants engage in an argumentative discussion seeking to persuade each other [2]. Understanding how people change their views is valuable in applications of social media analytics that involve modeling opinions and influence, such as in political debates, product evaluations, and diffusion of new ideas. Not surprisingly, online argumentative discussions are garnering increasing research attention [12, 19, 29].

Argumentative discussions on social media, such as Twitter, Reddit, and Quora, not only provide the basis for understanding people's thoughts and behaviors, but also enable the possibility of developing a customized information recommendation system [37, 38]. A fundamental concern for both participants and those who view or analyze discussions (e.g., for advertising or digital governance) is to deal with information overload in using knowledge from the web. Dealing with information overload requires a method that can effectively retrieve relevant information and plausible answers from massive online information.

Idiosyncrasies of individual users under varying circumstances are a source of complexity. Despite increasing attention on argumentation in social media, how and when view changes happen during an exchange of comments lacks investigation. Previous works about persuasiveness focus on statistics such as length or branching factors of the discussions, linguistic features of individual comments, and topic similarity between paired comments. In contrast, we posit that it is necessary to examine a discussion as a whole to capture the *interplay* of the exchanges between participants and their *cumulative* effect on the mental state of a participant.

We investigate how to use the most of the information in an online argumentative discussion from these two aspects. Table 1 gives a motivating example drawn from Reddit's Change My View forum (with names changed) involving one opinion holder (Alice) and multiple challengers who hold the opposite view (Bob, Chris, Dave, and Ella).

Table 1's scenario is a discussion initiated by Alice about the game mechanic of having random encounters in a game. Challengers raise three points in opposition: (1) old games use random encounters, (2) users prefer random encounters, and (3) random encounters are addictive. However, the second point ⟨PT 2.0⟩ raised by Bob is ignored by Alice at first. Chris and Dave argue with Alice on the third point, which though not explicitly related to ⟨PT 2.0⟩ builds up an effect on Alice about how random encounters make the game addictive and fun. And, when Ella observes in ⟨PT 2.1⟩ that some people love the game mechanic, Alice changes her view stating "I suddenly got sentimental about the idea." Thus, ⟨PT 2.0⟩ and ⟨PT 2.1⟩ raise the same point but receive different credit, indicating that what matters is not the effect of one individual comment, but the *cumulative influence* building up unconsciously via a series of comments—⟨PT 2.0⟩, ⟨PT 3.0⟩, and ⟨PT 3.2⟩ leading to ⟨PT 2.1⟩.

Such scenarios motivate us to focus on cumulative influence. We approach the problem of predicting *when* in a discussion an opinion

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380302

Table 1: A snippet of an online argumentative discussion. *OH* is an opinion holder and *C* is a challenger. Index indicates the threading structure (e.g., comment 1.1 is a child of comment 1). Three styles of underlining delineate three opposing points.

Index	Person	Role	Comment
1	Alice	OH	Random encounters, like in the Final Fantasy series, are a bad game mechanic.
1.1	Bob	C	I'd argue that it's not bad per-se, but just outdated and has no place in current games. ⟨pt 1.0⟩ The only reason it exists is because old gaming systems couldn't handle anything else at the time, but ⟨pt 2.0⟩ it persisted through generations because people liked it.
1.1.1	Alice	OH	⟨pt 1.1⟩ The part about old gaming systems not being able to handle anything else isn't true.
1.2	Chris	C	⟨pt 3.0⟩ Random reinforcement is a well-known way to make things more addictive
1.2.1	Alice	OH	⟨pt 3.1⟩ Being more addictive doesn't make it a good game mechanic.
1.2.1.1	Dave	C	⟨pt 3.2⟩ Good or bad or fun is subjective with regards to games.
1.3	Ella	C	⟨pt 2.1⟩ A number of people who like gambling note that they enjoy that sort of mechanic.
1.3.1	Alice	OH	⟨pt 2.2⟩ I suddenly got sentimental about the idea of random encounters and not quite wanting them to go away.

holder's view changes by modeling the interplay between participants' comments in the context of the discussion in a sequential manner. Therefore, we identify three research questions:

RQ Feature Is modeling the interplay of comments beneficial (and if so how much) in predicting an opinion holder's view change?

RQ Structure What representation of the sequential context helps predict view changes effectively?

RQ Benefit How does it help in practice to predict view change in the context of a whole discussion?

Our novelty lies in two aspects. First, we introduce the problem of predicting view changes in a more realistic setting—predicting when view change occurs—than previous approaches, which merely compare comments. Our setting respects the natural imbalance in our data where only a few of a large number of comments change anyone's view. Predicting persuasiveness in this setting is challenging because of idiosyncrasies of participants, changing focus (theme) during a discussion, and open-ended user-generated contents. Second, we make the first attempt of modeling cumulative influence on opinion holders when predicting the persuasiveness of arguments. During an online discussion, changes of mental state are reflected in the participant's comments with time. Therefore, we identify and use information about the context and the interaction information in our model to represent the cumulative influence.

To this end, we adopt a sequential model using compound features to capture the cumulative influence of a whole discussion. Our results show that a sequential model outperforms the competing nonsequential model by 4% in AUC-ROC (Area Under Curve for its Receiver Operating Characteristic) and by 9% in AP (Average Precision). The results support our claim that considering the cumulative influence and interplay of all participants' comments is important in predicting view changes in a discussion.

2 DATASETS

We consider two datasets: threaded discussions on Reddit Change My View (CMV) subreddit, and from human studies.

2.1 Reddit Change My View (CMV) data

On Reddit's CMV forum, opinion holders award a *delta* (which we write as Δ, for readability) to the comments that changed their view to some extent. Each discussion is terminated after three hours. We categorize CMV posts into three types, based on where they fit in a discussion and examine them to motivate our features.

- Original post (OP)—the opening by an opinion holder (OH), which provides the initial prompt including a justification of the view in a minimum of 500 characters.
- Comment from a challenger—a reply to the OP or to any subsequent comments. Typically, challengers refer to previous posts to provide counter-arguments. Agreement with the OP is prohibited in replies by CMV rules.
- Reply from the OH—to a comment from a challenger. The OHs may award Δs to challengers who caused changes to their views. We use Δs as our ground truth. However, there is no way to quantify the degree of the view change or force a Δ award by the platform.

We use the dataset provided by Tan et al. [34]. The raw dataset contains all discussions from January 2013 till August 2015, including both textual content and information such as user names, user scores, and timestamps. We filter discussions using criteria similar to Tan et al.'s. First, the fact that opinion holders enter the discussion asking for their views to be changed does not always indicate an open mind—only fewer than half of the discussions result in a Δ being awarded, and only a small portion of the comments (≈2%) wins a Δ. Second, we exclude discussions where the OP never replied, no Δ was awarded by the OP, or there were fewer than 10 replies. Table 2 shows the statistics of our dataset.

Table 2: Statistics of the CMV dataset. Here, cmt. and dis. stand for comments and discussions, respectively.

Count of	cmt.	cmt. w/ Δ	dis.	words per cmt.	sentences per cmt.
Training	447,788	7,620	5,000	96.9	5.3
Testing	59,161	1,157	683	96.7	5.2

Our observation corroborates previous work [34] that it is difficult to identify comments that earned Δs due to the idiosyncratic ratings by opinion holders in CMV. Therefore, we conduct empirical user studies to, first, establish the validity of the data in terms of the reliability of assignments of labels and, second, better understand the nuances of persuasive comments given their contextual and interactive features.

One note is that both Reddit CMV and the human study are settings with skewed demographically. However, this is the nature of many settings. For example, tweets about different topics attract different demographic bases. Although the feature set for this specific task may not be suitable in a different setting (i.e., features may have different weights or need to be extended leveraging domain-related factors), the results from the human study demonstrate that we need to put the model in the opinion holder's shoes to understand why a specific opinion holder's view would change or not change.

2.2 A Human Study

Due to the aforementioned fact that only a few of the comments win a Δ from an opinion holder, we conduct a human study on top of the CMV dataset in which participants estimate the soundness and persuasiveness of an argument. Our motivation is to provide additional finer-grained data to help us better understand the characteristics of online argumentative discussions.

We conducted the experiment with US military reservists as annotators. This study was approved by the Institutional Review Board (IRB) of our university and by the US Department of Defense. Informed consent was obtained from each participant. Considering task duration and expected qualities, we selected discussions with around 50 comments and randomly distributed them to annotators. Annotators were presented with comments from both opinion holders and challengers from a discussion and are requested to rate each comment as to its persuasiveness on a 1–5 scale, from weakly to strongly persuasive. Depending on time availability, each annotator rated one or more discussion threads.

We collected annotations of 72 discussions from 29 annotators. Table 3 shows the statistics of this dataset. Since we have a varying number of annotators for each thread, we calculated the inter-rater reliability using Krippendorff's alpha, obtaining 0.400, which indicates fair but not high reliability. The result suggests that judgments of persuasiveness are quite subjective. It is notable that the difference of average ratings between comments with Δ and comments without Δ is slight, whereas the differences of standard deviation and variance between Δ and non-Δ comments were substantial. This observation indicates that annotators, as observers rather than as participants (opinion holders), have the similar perception of persuasiveness of successful comments. The human study indicates that persuasive comments share certain traits, as discussed in previous studies [10, 35]. However, the large variance of ratings of comments without Δ indicates that perception of persuasiveness varies with each individual. To retrieve a comment that is both persuasive and effective for a certain opinion holder, we need to consider factors related to an opinion holder's perception. We present further analysis and additional discussions on the human study data in Section 5.1.

With this human study, we enrich CMV data by annotating persuasiveness ratings on a scale at the comment level and from multiple annotators. This provides additional information for understanding individual perception of persuasiveness and modeling how people change their views.

Table 3: Statistics of the human annotated data. Here, r denotes ratings. avg, std, and var stand for average, standard deviation, and variance, respectively.

Comments	count	$avg(r)$	$std(r)$	$var(r)$
Without Δ	2844	3.177	1.042	1.567
With Δ	102	3.419	0.703	0.631

3 APPROACH

We formalize our problem as a sequence labeling task. Each discussion in CMV is represented as a sequence of comments sorted in increasing order of time. Our objective is to predict whether a comment in the sequence wins a Δ. Our model's input is a sequence of comments $\boldsymbol{x} = (x_1, \ldots, x_n)$, each with a set of feature values. The model's output is a sequence of binary labels $\boldsymbol{y} = (y_1, \ldots, y_n)$ for each comment in $\boldsymbol{x}$. A positive label (i.e., 1) means the comment changes the view of the opinion holder and wins a Δ; a negative label (i.e., 0) means it does not.

An overview of the proposed supervised sequential model (CRF) is shown in Figure 1. Discussions are encoded offline using identified features as discussed in Section 3.1. We discuss how to use CRF to represent cumulative influence and how to generate sequences from threaded discussions in Section 3.2 and Section 3.3.

3.1 Features and Revealed Traits

We coarsely categorize features influencing view change as linguistic, contextual, and interactive. This categorization is extensible since each category represent a meaningful aspect: linguistic features represent characteristics of individual comments and can be extracted purely from texts, contextual features are metadata (dynamics) of a discussion as are provided by the platform API or can be obtained by calculating statistics, and interactive features represent the pairwise relations between comments.

Our selected feature set reflects the traits of CMV comments in terms of the usage of a language, the position of a comment in the context of a discussion, and the interplay of comments.

We describe the methods for computing and quantifying features in detail in Section 4.2. We don't seek to explore an exhaustive list of features but to demonstrate features that work well for predicting persuasiveness. However, our model is general enough to incorporate additional features that may be available for specific problems. The linguistic and contextual features are selected from the relevant literature in psychology and communication theory. Our key novelty lies in representing interactive features in a sequential model.

Linguistic features reflect persuasion strategies adopted by challengers. We employ linguistic features that are known to correlate with topics [11], argumentation, personality, and persuasion success. Linguistic features are extracted from the text of each comment with a lexicon-based approach.

Definite and indefinite articles often express specificity and generality, respectively. We might expect that people would prefer more specific over less specific information.

First-person and second-person pronouns are indicators in aspects such as openness, deception, and persuasiveness [9, 25, 28].

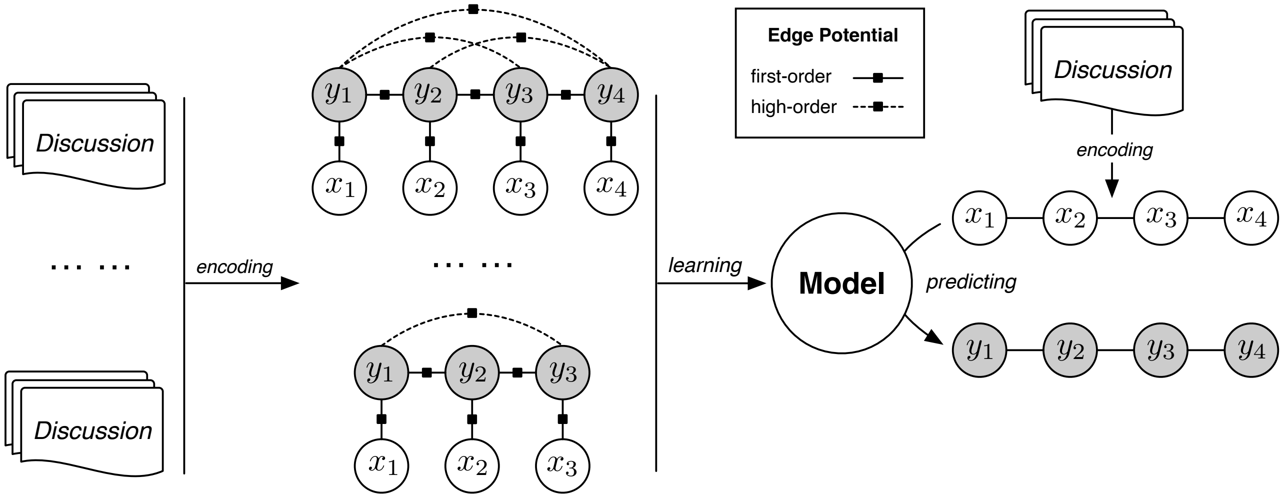


Figure 1: Overview of the proposed CRF model.

A first-person pronoun indicates individual experience, personal judgment, or affirmation of the information conveyed.

Hedge words indicate uncertainty, which can indicate a weak statement or rationality and prudence. In addition, hedging is associated with indirection in politeness theory [4], which may have an effect on an opinion holder's mental state.

Sentiment strongly indicates subjectivity by conveying the polarity of an opinion. It proves effective in social network link prediction [39] and precision advertising.

Biased language may indicate a weakness of one's argument. We employ linguistic cues of two major classes of bias—framing (subjectivity) and epistemological (believability of a proposition) [31].

Other features include examples, questions, and links, which prior work ranks as important to persuasion [29, 34].

Contextual features include those pertaining to discussions at large as well as to users. Contextual features reflect the dynamics of a discussion and the situation of a certain comment. Contextual features are extracted based on properties of the discussion. CMV provides information such as timestamps of posts and user's contribution scores.

Length of discussion indicates the OH being entrenched and the topic being controversial.

Entry order of a comment is an indicator of cumulative influence up to that comment. A challenger entering late in a discussion may absorb previous information and gain a better understanding of the opinion holder's reasoning before commenting. However, previous studies [34, 38] show that an earlier comment is more likely to earn a Δ than a later comment, possibly due to a first-mover advantage.

Authorship identifies OH and challengers. It reflects which parties the interplay is between.

User's karma score and historical Δ*s* indicate a user's past engagement, especially historical discussions and contributions. Karma is calculated as the number of upvotes minus the number of downvotes earned by the user.

Interactive features capture explicit interaction (replies to and quotations from comments) as well as implicit interaction (relationships such as discourse structure and persuasion strategy). We examine implicit interactions at a coarse level that capture the existence of potential connections of two comments but not the type of a certain connection.

Direct response to a comment indicates turn taking by the OH and the challengers, sometimes ending with a Δ being awarded. Direct responses indicate coherence of two comments.

Quotation reflects that the points made in a comment repeat what was presented before but presumably with additional evidence or reasoning.

Connection with OP indicates an understanding of OH's reasoning in the original post. Such connections appear as topic relevance. We measure relevance as text similarity and describe the computation in detail in Section 4.2.

Connection with OH's latest reply indicate whether a challenger engages in a lively discussion. The connection is measured in the same way as with OP.

Response from OH reflects an OH's opinion on a counter-argument. For example, an OH who has strongly resisted a point in a counter-argument may not be persuaded by the same point again. Therefore, we identify dependencies between two comments and use the OH's sentiment to the previous comment as a feature to predict the label of the later comment.

3.2 Sequential Model

We adopt Conditional Random Fields (CRF) [20] to capture both adjacent and nonlocal dependencies between comments.

Given a sequence of vectorized comments $\boldsymbol{x}$, our model predicts a sequence of corresponding labels $\boldsymbol{y}$ for each comment in $\boldsymbol{x}$ using conditional probability $P(\boldsymbol{y}|\boldsymbol{x})$. In general, vertices V of a CRF graph G are partitioned into cliques $V_G = \{C_1, \ldots, C_p\}$. Given a set of features, θ represents parameters of a CRF model to be estimated.

$P(y|x)$ with parameters θ is written as in Equation 1.

$$P_\theta(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(x)} \prod_{C_i \in V_G} \prod_{\Psi_c \in C_i} \Psi_c(\boldsymbol{x_c}, \boldsymbol{y_c}) \tag{1}$$

Here, each edge factor Ψ_c represents an edge in a clique c as shown in Equation 2: $Z(\boldsymbol{x})$ is a normalization factor; f_k is a feature function; λ_k represents the weight for each feature. Parameters $\theta = \{\lambda_k\}$ are estimated to maximize log-likelihood during the training phase.

$$\Psi(\boldsymbol{x_c}, \boldsymbol{y_c}) = \exp \sum_k \lambda_k f_k(\boldsymbol{x_c}, \boldsymbol{y_c}) \tag{2}$$

Linear CRF. When the graph structure is a sequence, the conditional probability follows the Markov assumption and considers only $\Psi(y_t, y_{t-1}, \boldsymbol{x})$. Linear CRF is the common type of CRF in NLP, such as for POS tagging, sentence segmentation [40], and sentiment flow modeling [22].

Skip-Chain CRF. Skip-chain CRF uses a general CRF graph representation with higher-order dependencies. In a skip-chain CRF model, nonadjacent but dependent nodes are linked by a skip edge and such long-distance dependencies are included in the conditional distribution. Specifically, skip-chain CRFs capture nonlocal dependencies between pairs of nodes. The skip-chain CRF combines two kinds of dependencies: (1) first-order of the form $\Psi(y_t, y_{t-1}, x)$ with first-order feature functions $\{f_1, \ldots, f_K\}$, as in linear CRF (Equation 3 and edges in solid lines in Figure 1) and (2) higher-order of the form $\Psi(y_u, y_v, x)$ with higher-order feature functions $\{f_{K+1}, \ldots, f_J\}$ where $|u - v| > 1$ (Equation 4 and edges in dotted lines in Figure 1).

$$\Psi(y_t, y_{t-1}, \boldsymbol{x}) = \exp \sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \boldsymbol{x}) \tag{3}$$

$$\Phi(y_u, y_v, \boldsymbol{x}) = \exp \sum_{k=K+1}^{J} \lambda_k f_k(y_u, y_v, \boldsymbol{x}) \tag{4}$$

3.3 Sequence Structures

A challenge in modeling discussions with CRF is to capture dependencies (i.e., skip edges) and context (i.e., edge factors) properly.

Capture nonlocal dependencies and context. As Table 1 illustrates, we identify a pair of dependent nonadjacent comments based on the interactive features. We construct a skip edge between two comments if (1) one comment is the child of the other or (2) the textual similarity between the two comments is above a certain threshold. We term these criteria Skip-Rules. Algorithm 1 describes how to form a sequence and skip edges for CRF with Skip-Rules.

Form subsequences. Another consideration is the size of the discussions. In our dataset, the longest discussion contains 1,980 comments and 88 branches. With additional edges representing dependencies for content-based similarities, the number of nodes in the maximum clique of the graph is large and inference can be quite inefficient. Therefore, we break a discussion down into subsequences bounded by the opinion holder's participation (i.e., a reply from the opinion holder to previous comments).

Algorithm 1: Form sequences and skip-edges with dependencies and context.

```
Input: A sequence of comments sorted by time
       $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$
Output: A sequence of comments encoded in feature space,
        a list $E$ of node pairs that are linked by skip edges
Extract R a set of comments that are replied to by OH
for each $x_u$ in $\boldsymbol{x}$ do
    Extract linguistic and contextual features for $x_u$
    for each $x_v$ in $\boldsymbol{x}$ where $v < u$ do
        if $(x_u, x_v)$ satisfies a member of Skip-Rules then
            Add $(x_u, x_v)$ to $E$
            if $x_v \in R$ then
                $f(y_u, y_v)$ = sentiment score of OH's
                  response to $x_v$
            end
        end
    end
end
```

Subsequences structured as explained above respect turn-taking and model a discussion of multiple rounds. Each round is represented as a subsequence. It starts with a statement or reply by the OH followed by comments from challengers who take in the latest information present in the OH's replies.

The OH's mental state may change in each round based on information received from challengers. We assume the OH's mindset remains unchanged within a round but potentially changes across rounds and is reflected in the OH's reply at the beginning of each round. Accordingly, we construct each round as a subsequence.

4 EXPERIMENTS AND ANALYSIS

We describe the computational methods and configurations for the experiments in detail in this section. Our code and data are released at https://github.com/zguo8/cmv-cumulative-influence.

4.1 Data Preprocessing

As a prelude to building the feature space and forming skip edges, we preprocess the data to better represent specific features. From the original dataset [34], we filter out discussions (1) in which an opinion holder never replied or awarded a Δ and (2) that have fewer than 10 comments. We also exclude the "initiating" period of CMV (before May 2013), which shows a different level of counts of replies and counts of participants from the later period.

We remove non-ASCII characters, fill periods in place of line breaks if punctuation is missing, and exclude system messages such as deleted message marks.

4.2 Feature Extraction

Table 4 describes the methods used for quantifying the features and the lexicons used in the computation.

For measuring text similarity between comments, we tokenize a comment into sentences, encode the sentences using the Universal Sentence Encoder [6], and calculate the cosine similarity for each pair of sentences in the two comments.

Table 4: List of features and explanations. Here, $len(d)$ denotes the length of a discussion, $len(c)$ denotes the length of a comment, and $len(subseq)$ denotes the length of a subsequence measured by the count of comments in the subsequence.

Categories	Features	Explanation
Linguistic	Definite/indefinite articles	Occurrences normalized by $len(c)$
Linguistic	1st/2nd person pronouns	Occurrences normalized by $len(c)$
Linguistic	Hedges	Use the list of hedge words created by Hyland [18]
Linguistic	Sentiment	VADER compound scores [17]
Linguistic	Biased language	Occurrences of each subtype of biased text [31]
Linguistic	Examples	Occurrences of "for example" and alternative expressions
Linguistic	Questions	Count of question marks
Linguistic	Links	Count of "http" and "https" marks
Contextual	Length of a discussion	The count of the comments in a discussion.
Contextual	Entry order of a comment	Normalized by $len(d)$ in settings with sequences; normalized by $len(subseq)$ in settings with subsequences
Contextual	Authentication	A binary value where 1 denotes an opinion holder and 0 denotes a challenger
Contextual	User's flair	A score on Reddit indicating a user's contribution provided by the Reddit API
Interactive	Direct response to	Use the index of the precedent comment to which the current comment replies
Interactive	Quotation	"&gt;" marks in Reddit comments
Interactive	Connection with the OP	Measured by cosine similarity with USE (Universal Sentence Encoder) [6]
Interactive	Connection with OH's most recent replies	Measured by cosine similarity to the latest reply of OH; Measured by cosine similarity with USE embeddings
Interactive	Response from OH	OH's sentiment to the previous most-similar comment (measured by cosine similarity with USE embeddings)

4.3 Model Development and Configuration

We realize the linear CRF model using the sklearn-crfsuite implementation, which is an open-source Python wrapper for CRFsuite [26]. We inherit default settings of sklearn-crfsuite, where parameter estimation is performed by the L-BFGS algorithm with elastic net (L1 and L2) regularization. For labeling each item in a sequence, the library uses Viterbi algorithm and calculates marginal probabilities to find the most likely class label.

To capture higher-order dependencies, we develop a model with the general graph CRF implementation with PyStruct [23]. The training data of the model contains two parts. One is the graph structure and the other is the feature space. During the training phase, we use the default structured SVM solver to estimate the margin as a convex optimization problem. When predicting the items in a new sequence, it approximates the maximum a posterior probability (MAP) for labeling. For the skip-chain CRF, we choose linear programming as the inference algorithms to obtain marginal probabilities to plot the curves.

4.4 Metrics

Since our data are highly imbalanced, we choose the following metrics to evaluate the performance of our model. Section 5 discusses how we apply these metrics.

AUC-ROC. AUC (Area Under the Curve) scores of ROC (Receiver Operating Characteristic) measure the true positive rate (TPR) against the false positive rate (FPR). AUC score of ROC reflects whether a comment that received a Δ is predicted to have a higher probability of a positive class label than a comment that did not receive a Δ. This metric is applied in predicting persuasiveness of arguments [19] and malleability of initial statements [34].

AP. Average Precision (AP) is calculated as the area under the Precision Recall Curve (PRC). PRC has gained attention in evaluating imbalanced data, since ROC provides a misleading interpretation and visualization of specificity (i.e., true negatives among predicted negatives) on imbalanced data [30, 32]. PRC provides an interpretation of specificity complementing AUC-ROC for evaluating the performance of CRFs on imbalanced data [3, 16].

MRR. Mean Reciprocal Rank (MRR) is the mean across all discussions of the reciprocals of the ranks at which the first relevant item (i.e., a Δ-earning comment) is retrieved.

4.5 Baselines

Since our specific problem—predicting the points of view changes in the context of an overall discussion—is novel, there is no previous work we can directly compare to. For example, Tan et al. [34] tackle a binary classification task where input data are paired root replies—such corpus does not provide the feasibility to examine the cumulative influence from the interplay of other comments. Therefore, we evaluate our model from multiple angles as follows.

Nonsequential models. We choose LR (logistic regression) and SVM (Support Vector Machine) as our nonsequential baseline models. LR is a strong baseline, as it is employed in the aforementioned related work [19, 29, 34]. SVM is powerful for NLP tasks including determining the persuasiveness of an argument [7]. For logistic regression, we explore L1 and L2 regularization with different regularization strengths. We adopt the SVM model in scikit-learn [27],

	Linear CRF		LR		SVM	
	OUR	Tan	OUR	Tan	OUR	Tan
AP	**0.23**	0.19	**0.18**	0.07	**0.09**	0.07
AUC	**0.87**	0.79	**0.86**	0.77	**0.81**	0.77
MRR	**0.51**	0.49	0.48	**0.49**	**0.40**	0.35

(a) Feature gains across models.

	Linear CRF			Skip-chain	Random
	Seq	Subseq	RepCmt	CRF Seq	
AP	0.23	**0.27**	0.31**	0.14	0.02
AUC	0.87	**0.90**	0.73	0.70	0.50
MRR	**0.51**	0.78*	0.58**	**0.51**	0.11

(b) Performance across models and structures.

Table 5: Comparing performance of selected models. OUR = Our feature set. Tan = Tan's interplay features. LR = Logistic Regression. RepCmt = examined on only comments replied by OH. * MRR is highest (0.78) in Linear CRF with Subsequences structure and our feature set, but it ranks over subsequences (≈ 10 comments on average) as stated in Section 3.3 whereas others rank over discussions (≈ 90 comments on average). ** Similarly, MRR and AP of RepCmt are not comparable since they are calculated only over comments that an OH responds to (≈ 10 comments on average).

in which the probabilities are calibrated using Platt scaling with additional cross-validation. For SVM, we explore linear and radial basis function (RBF) kernels. For both baselines, the class weights parameter is set to 1:5 respecting the imbalanced data. The chosen weight (1:5) performs best among a series of weights in increments of 5, specifically, 1:10, 1:15, 1:20. With this setup, the models set a smaller penalty for false positives than for false negatives.

Tan et al.'s interplay features. To evaluate the identified contextual and interactive features, we compare our feature set to Tan et al.'s interplay features by applying them on each model. Tan et al. [34] evaluate multiple feature categories such as the number of words, BOW, POS, interplay, and style features and find that the interplay features significantly outperform other features and are more robust when controlling for the length of comments in experiments. Therefore, we implemented Tan's interplay features, encoding the comments use these features, and compare performance of a model with our proposed features to the same model but with Tan's interplay features.

Comments replied by OH. Jo et al. [19] investigate interactions through a neural network by pairing each sentence in a challenger's comment with every sentence in the OP. However, they examine *only* comments replied to by the OH and omitted comments that the OH intentionally or unintentionally ignored. To compare within Jo et al.'s result, for this specific evaluation, we only use comments that OH replied to as our input rather than all comments.

Human annotation. We obtained the ratings from reservists on a sample dataset of CMV discussions. These human annotators presumably hold different prior beliefs and have different judgments about persuasiveness from the opinion holders. Comparing to the models, the judgment from human annotators is subject to cognitive process, thus is not limited by identified features. Note that for the human annotation task, we select the discussions whose number of comments is around 50. Therefore, it results in a test set with a different positive rate (3%), which is higher than the whole original CMV dataset (2%). As an additional experiment, we test and report the results of our models for this specific dataset and compare with human performance in Section 5.1.

Random guessing. Tan et al. [34] conduct an experiment about malleability of original posts from opinion holders. They asked annotators to find the malleable one from a pair of original posts. Tan et al. report that human annotators performed no better than random guessing and LR performs only slightly better than random guessing on predicting malleability. Although our problem setting is different from the malleability study, the comments of CMV discussions yield the same randomness and subjectivity regarding their persuasiveness to opinion holders. Therefore, we include random guessing as one of the baselines. Random guessing has an AUC-ROC of 0.5 independent of the class distribution. However, the AP for random guessing is determined by the positive rate, which in our setting, is 0.02. MRR of random guessing for the test set is 0.11, which is calculated by $\frac{1}{count(d)} \sum_{d} MRR(d)$, where $MRR(t) = \frac{1}{len(d)} \sum_{r=1}^{len(d)} \frac{1}{r}$.

5 ANSWERING OUR RESEARCH QUESTIONS

We now revisit the research questions motivated in Section 1. Table 5 shows the comparative results—subsequent to hyperparameter optimization and three-fold cross-validation for each model.

RQ Feature. For LR, our feature set outperforms Tan et al.'s (2016) interplay features by an absolute increase of 9% on AUC-ROC and 11% on AP. Linear CRF with our feature set outperforms that with Tan et al.'s feature set by 8% on AUC-ROC and 4% on AP.

Table 6 shows the performance of each feature category using CRF. It indicates that contextual and interactive features improve performance, which supports our claim that considering the cumulative influence of all participants' comments is important for predicting view changes.

Table 6: Scores for feature categories.

	Linguistic	Contextual	Interactive	All
AP	0.13	0.19	0.17	0.23
AUC	0.73	0.84	0.81	0.87

To take a close look at the impact of each feature, we examine the feature weights in linear CRF. The coefficients of normalized feature values in linear CRF are learned using the optimization algorithm and represent the contribution of each feature to class prediction. Top-ranking features with positive and negative contributions to linear CRF are listed in Table 7.

Table 7: Top correlated features from Table 4.

Positively Correlated		Negatively Correlated	
Feature	Coefficient	Feature	Coefficient
Links	0.117	relative order	−0.674
2nd person pron.	0.008	Sim. w/ recent	−0.064
User flair	0.005	Sim. w/ OH	−0.055
Definite articles	0.003	Quotation	−0.040
Indefinite articles	0.001	Questions	−0.033

RQ Structure. With our feature set, linear CRF with subsequences outperforms linear CRF with sequences by an absolute increase of 4% in AUC and of 3% in AP. The results illustrate that modeling a CMV discussion with subsequences bounded by the OH's participation has practical significance in predicting that OH's view change, as discussed in Section 3.3.

Although skip-chain CRFs capture more dependencies than linear CRFs, they perform worse. A potential explanation is that the structure is given a priori, not learned. We evaluate additional potential long-range dependencies such as linking comments from the same challenger with skip edges. However, additional edges result in dramatically increased computing time but no improvement in the prediction. Because of the intractability of approximate inference of skip-chain CRFs and the complexity of confounding indicators of dependencies, it is not practical to improve the performance of skip-chain CRFs with hand-crafted rules. We defer to future work to integrate additional layers such as recurrent neural networks to learn dependencies.

One may ask how general CRFs that represent the full-threaded tree structure of a discussion would fare. We conduct an additional test on a tree-structured CRF comparing to skip-chain CRF with different rules for constructing skip edges. Table 8 shows how various rules for constructing edges affect performance. Although tree structure achieves higher AUC-ROC, skip-chain CRF with linked similar comments achieves higher AP. This finding suggests that tree structure predicts more true positives though at the cost of a larger number of false positives than does skip-chain CRF.

Table 8: Performance comparison when constructing skip edges based on different rules.

	Tree	Reply to	Identity	Similarity
AP	0.09	0.06	0.03	**0.14**
AUC	**0.78**	0.59	0.50	0.70

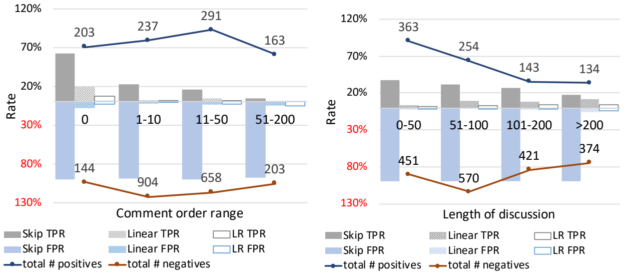
RQ Benefit. Predicting view changes in real life is challenging because OH's propensity to award deltas is an unknown prior and data are highly skewed due to the idiosyncrasies of individual OHs. A model that works for pairwise comments with a balanced dataset would not be suitable because real-life data is far from balanced. The proposed sequential model does not require balanced data, and is thus applies better in a natural setting.

An obstacle when evaluating the sequential model on the highly skewed CMV data is that there is no other indicator whether non-Δ comments are persuasive to some extent. Accordingly, examining the top candidates (predicted positives with high possibility or confidence) rather than exact hits of positives is appropriate. Therefore, we adopt a user model where a user queries for points when an OH's view may plausibly change during a discussion. We adopt this user model from information retrieval tasks, specifically, of results produced by a search engine in which the top passages matter most. In our setting, we care about if the model can place the most likely comments among the top candidates.

The reciprocal of MRR indicates how many comments a user needs to examine before finding the right one. On CMV, a user needs to read 50 comments on average to find the most persuasive comment if the discussion contains 100 comments. However, an MRR (0.51) for linear CRF with sequences (i.e., an overall discussion) indicates that the model can rank the most persuasive comment among the top two on average.

We conduct statistical tests for better understanding the performance difference between the sequential model and the baseline nonsequential model. Due to the nature of the CMV data, where no assumption may be made about the distribution, we use McNemar's test to evaluate the significance for which $p < 0.05$ indicates a statistically significant difference between evaluated models and Cohen's g index (also known as Cohen's d_s) to evaluate the effect size for which < 0.15 indicates a small effect size, $0.15 - 0.25$ indicates a medium effect size, and > 0.25 indicates a large effect size [8]. We compare prediction results from LR and linear CRF with subsequences on paired samples. We test on prediction results in the test set and obtain a p-value < 0.001 with McNemar's test and Cohen's g index of 0.384. The p-value indicates a rejection of the null hypothesis and the g index indicates a large effect size, i.e., there is a significant difference between the predictions from LR and the predictions from CRF.

To supplement this result, we analyze the true and false positive rates of the models against discussion dynamics. Figure 2 shows that linear and skip-chain CRFs identify more true positives than logistic regression for longer discussions and for plausible comments that come later in a discussion.



(a) TPR/FPR against time order of a comment. **(b) TPR/FPR against discussion length (# comments).**

Figure 2: True Positive Rate (TPR) and False Positive Rate (FPR) with respect to discussion dynamics.

5.1 Lessons Learned from Human Study

The above analysis indicates that sequential models demonstrate their ability to capture the cumulative influence and individual perception to some extent. However, the average precision is relatively low comparing to other information retrieval tasks. Therefore, we took a closer look at the human study results to shed some light on the factors that need further investigation in future research.

The human study provides a dataset at a finer granularity than CMV since each comment receives a rating on a scale from multiple annotators rather than a binary label solely from an opinion holder. We obtain a Pearson's correlation coefficient for raters' average ratings and binary Δ labels of 0.1645. This weak correlation is consistent with the large variance of raters' ratings on the same comment and indicates that idiosyncrasies of an individual user greatly affect his or her perception of persuasiveness.

To directly compare the performance of our model to human annotations, we evaluate the best performing CRF model on the human study data and evaluate human annotations by using average ratings as a proxy of the probability for a comment to receive a Δ. As shown in Table 9, the CRF model achieves better performance on all metrics than human annotations. It is notable that human annotations achieve a low AP value similar to the nonsequential models, LR and SVM. In this study, we observe that the low AP value from humans arise primarily because annotators give a greater number of "is-persuasive" ratings (4 or 5) than the actual number of comments with Δs.

Table 9: Performance comparison between human annotations and linear CRF.

	AP	AUC	MRR
Human	0.09	0.77	0.45
CRF	0.23	0.85	0.81

On the one hand, low average precision is natural for this challenging task that attempts to retrieve the most effective information to change someone's view from the massive user-generated contents. On the other hand, the results reveal bias in individual perceptions of information. For example, human annotators tend to give high ratings to long comments, as shown in Figure 3.

In addition, we test how our trained CRF model performs if we take human annotations as ground truth for testing. The best performing linear CRF model achieves AUC-ROC of 0.78, AP of 0.32, and MRR of 0.24. The AP is higher than results on the original larger test set, since there are more positives in this ground truth. The experimental results and these findings indicate that the weight of each feature may vary across different settings and the model should be adjusted (retrained) accordingly.

6 UNDERSTANDING PERSUASIVENESS

To overcome the prior belief and bias from researchers, we conduct an additional study (determined to be exempt by our IRB office) to understand the reasoning behind the variance of individual perception of persuasiveness. The study involved 46 participants who are students majoring in computer science. Each participant was randomly assigned two CMV discussion threads. For a selected subset of the comments from its two threads, a participant was asked to rate the persuasiveness of each comment and provide a justification for each rating. Seeing potential reasons helps us investigate why our model yielded a low average precision. Therefore, we select comments for collecting annotators' reasons if a comment satisfies one of the following criteria: (1) the comment received a Δ from an opinion holder but the CRF model predicted it as not persuasive (false negatives with low possibilities); or (2) the comment did not receive a Δ from an opinion holder but the CRF model predicted it as persuasive (false positives with high possibilities).

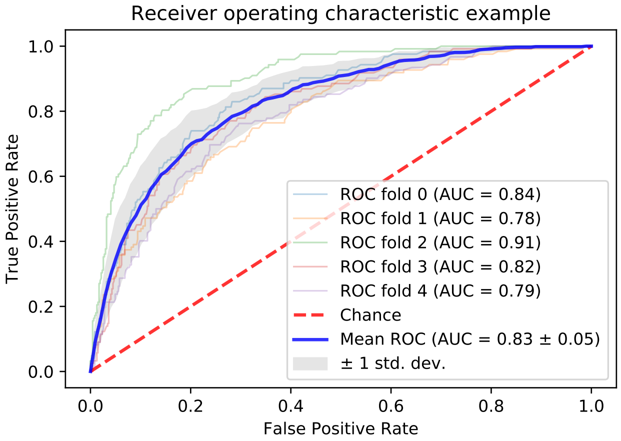


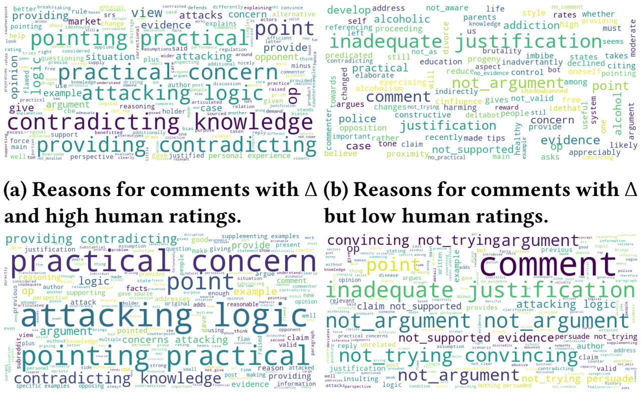
Figure 3: AUC-ROC of human ratings using number of words as a feature.

We collected 590 reasons for 32 CMV discussions. Since annotators' ratings do not always align with opinion holders', we divided the reasons into four groups: (1) 76 comments with Δ and high human ratings, (2) 28 comments with Δ but low human ratings, (3) 249 comments with no Δ but high human ratings, and (4) 237 comments with no Δ and low human ratings. Bigrams that arise frequently in these reasons are shown as word clouds in Figure 4.

The collected reasons from this study provide insights on persuasiveness of comments beyond language usage. The most frequently mentioned reasons for persuasive comments, as shown in Figures 4a and 4c, are providing contradicting knowledge, practical concerns, attacking logic, and examples. It is notable that attacking logic is a frequent strategy that is accepted by human annotators but refused by opinion holders (Figure 4c). For comments that receive low ratings, annotators provide reasons such as inadequate justification, lack of evidence, and not trying to be convincing. Figure 4b suggests that despite the flaws identified by annotators in a comment, opinion holders may find the comment persuasive. Figure 4d indicates features such as the quality of arguments (e.g., detecting a lack of evidence) that a model should incorporate.

7 RELATED WORK

Predicting view change in online argumentation is difficult because it involves cumulative influence, language factors, personality differences, and other confounding factors. Prior research approaches the problem from three main angles.



(a) Reasons for comments with Δ and high human ratings.

(b) Reasons for comments with Δ but low human ratings.

(c) Reasons for comments with no Δ but high human ratings.

(d) Reasons for comments with no Δ and low human ratings.

Figure 4: Word clouds showing the frequent reasons for high and low ratings of persuasiveness.

The first angle is studying the persuasiveness of an individual comment or an individual participant. Tan et al. [34] study a binary classification task of predicting persuasiveness where two similar comments are given, which forms a perfectly balanced dataset. Tan et al. also study the language factor of only original posts from OHs to predict whether or not an OH will change his or her view. However, this work does not consider (1) the interplay between challengers' comments and OH's replies and (2) dependencies between features in successive comments. Likewise, Jo et al. [19] study the relations between topics and malleability; they consider the interplay of the original post and a counter-argument but not in the context of the whole discussion. Villata et al. [36] study persuasiveness of arguments and engagement of participants in small groups; however, their study focuses on one participant rather than mutual effects across participants.

A second angle is considering the contextual factors as discussion dynamics and characteristics. For instance, Lukin et al. [21] consider audience factors (i.e., OH's personality here), which they capture via surveys before a discussion. However, surveys are not an effective method in online discussions. Tan et al. [34] find that the entry order of comments is predictive of their persuasiveness. Wei et al. [38] explore the dynamics of the reply tree of a discussion and users' reputations. Both works investigate correlations between persuasiveness and each factor separately. However, we posit that factors influence each other and should be estimated jointly. Therefore, we propose a sequential model that use compound features representing the cumulative influence of a whole discussion from the following aspects: (1) we use the sentiments of an OH's responses to capture a shift in his or her mindset over time; (2) we encode comments using predictive factors and use the encoded vectors to determine the existence of dependencies between successive and nonsuccessive comments; (3) we represent dependencies as conditional probabilities and examine their cumulative influence on an opinion holder's mindset.

The third angle focuses on argumentation mining. Analyzing the persuasiveness of arguments is an important direction within argumentation mining. Studies on argumentation mining analyze argument components via formal schemes to analyze the quality and logical connections among the components [5]. For instance, Hidey et al. [14] study claims and premises, and [29] study motions, assertions, and justifications. The associated annotations of persuasiveness or solidness are, in principle, objective [12, 13]. However, in contrast with annotators who may be instructed to reduce bias or be objective, opinion holders on social media are often biased. Thus, we model contextual information and dependencies of comments from a different angle and defer applying knowledge from argumentation mining as future work.

8 CONCLUSIONS

This paper is the first work to raise the problem of identifying a specific point where a view change occurs in an argumentative discussion. This specific problem incorporates real-world factors such as idiosyncrasies of individuals under varying circumstances and information overload on the web. Our results show that modeling the interplay between comments with a sequential model outperforms nonsequential models.

Our approach can potentially be applied to any setting involving online natural language interactions—our emphasis on capturing the dynamic interplay between comments is a central concern in understanding social interaction in general. Future directions include detecting and incorporating argument relations in discussions and evaluating how them in new settings, such as the resolution of conflicts between stakeholders' goals in design tasks [1, 24].

Based on the findings of Section 6, we identify the following directions for future research.

Categorization of countering strategies. Top reasons, as shown in Figure 4a, indicate a need for classification of countering strategies and understanding the effectiveness of different strategies. A common method to categorize strategies for arguing is to use logos, ethos, and pathos [15]. However, this categorization is from the perspective of rhetorical appeals, and our study indicates that categorization in a finer granularity—which can distinguish contradicting knowledge from practical concerns—is necessary.

Detection of long-distance relations in conversations. Researchers have put effort on identifying relations among argumentation components in discourse [33]. Detecting relations of argument components in conversation is a similar task but requires additional effort on linking content from multiple participants. For instance, as shown in Figure 4b, some comments that earned Δs from opinion holders are marked as unpersuasive by human annotators with reasons "not an argument" or "just a comment." With a deeper investigation, we observe that these comments often relate to a claim in a previous comment and serve as an argumentation component. This type of long-distance relations is difficult to identify with linguistic indicators, since the transitional words are not stated and require external knowledge.

ACKNOWLEDGMENTS

Thanks to Jascha Swisher and Ken Thompson for valuable discussions, and the anonymous reviewers for helpful comments. Thanks to the NCSU Laboratory for Analytic Sciences for support.

REFERENCES

[1] Nirav Ajmeri, Chung-Wei Hang, Simon D. Parsons, and Munindar P. Singh. 2017. Aragorn: Eliciting and Maintaining Secure Service Policies. *IEEE Computer* 50, 12 (Dec. 2017), 50–58. https://doi.org/10.1109/MC.2017.4451210

[2] Amparo Elizabeth Cano Basave and Yulan He. 2016. A Study of the Impact of Persuasive Argumentation in Political Debates. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics, San Diego, California, 1405–1413.

[3] Andrew P. Bradley. 1997. The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition* 30, 7 (July 1997), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

[4] Susan E. Brennan and Justina O. Ohaeri. 1999. Why Do Electronic Conversations Seem Less Polite? The Costs and Benefits of Hedging. In *Proceedings of the International Joint Conference on Work Activities Coordination and Collaboration* (San Francisco, California, USA) *(WACC)*. ACM, New York, NY, USA, 227–235. https://doi.org/10.1145/295665.295942

[5] Elena Cabrio and Serena Villata. 2018. Five Years of Argument Mining: a Data-driven Analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, Stockholm, Sweden, 5427–5433. https://doi.org/10.24963/ijcai.2018/766

[6] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder. *CoRR* abs/1803.11175 (2018), 7. arXiv:1803.11175 http://arxiv.org/abs/1803.11175

[7] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology* 2 (2011), 27:1–27:27. Issue 3. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[8] Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates, Hillsdale, New Jersey.

[9] Ulla Connor. 1990. Linguistic/Rhetorical Measures for International Persuasive Student Writing. *Research in the Teaching of English* 24, 1 (1990), 67–87.

[10] Esin Durmus and Claire Cardie. 2019. Modeling the Factors of User Success in Online Debate. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. ACM, New York, NY, USA, 2701–2707. https://doi.org/10.1145/3308558.3313676

[11] Susan Ervin-Tripp. 1964. An Analysis of the Interaction of Language, Topic, and Listener. *American Anthropologist* 66, 6_PART2 (1964), 86–102.

[12] Ivan Habernal and Iryna Gurevych. 2016. What Makes a Convincing Argument? Empirical Analysis and Detecting Attributes of Convincingness in Web Argumentation. In *Proceedings of the 21st Conference on Empirical Methods in Natural Language Processing* (Austin, Texas). Association for Computational Linguistics, Austin, Texas, 1214–1223. https://doi.org/10.18653/v1/D16-1129

[13] Ivan Habernal and Iryna Gurevych. 2016. Which Argument is More Convincing? Analyzing and Predicting Convincingness of Web Arguments Using Bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Berlin, Germany). Association for Computational Linguistics, Berlin, Germany, 1589–1599. https://doi.org/10.18653/v1/P16-1150

[14] 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 4th Workshop on Argument Mining*. Association for Computational Linguistics, Copenhagen, Denmark, 11–21.

[15] Colin Higgins and Robyn Walker. 2012. Ethos, Logos, Pathos: Strategies of Persuasion in Social/Environmental Reports. *Accounting Forum* 36, 3 (2012), 194 – 208. https://doi.org/10.1016/j.accfor.2012.02.003 Analyzing the Quality, Meaning and Accountability of Organizational Communication.

[16] T. Ryan Hoens and Nitesh V. Chawla. 2013. Imbalanced Datasets: From Sampling to Classifiers. In *Imbalanced Learning: Foundations, Algorithms, and Applications*, Haibo He and Yunqian Ma (Eds.). John Wiley & Sons, Hoboken, New Jersey, 43–59. https://doi.org/10.1002/9781118646106.ch3

[17] Clayton J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media* (Sofia, Bulgaria). ICWSM, Ann Arbor, MI, 216–225.

[18] Ken Hyland. 2018. *Metadiscourse: Exploring Interaction in Writing*. Bloomsbury Publishing, London, UK.

[19] Yohan Jo, Shivani Poddar, Byungsoo Jeon, Qinlan Shen, Carolyn Rosé, and Graham Neubig. 2018. Attentive Interaction Model: Modeling Changes in View in Argumentation. In *Proceedings of the 16th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (New Orleans, Louisiana). Association for Computational Linguistics, New Orleans, Louisiana, 103–116. http://aclweb.org/anthology/N18-1010

[20] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 8th International Conference on Machine Learning (ICML)*. Morgan Kaufmann, Williamstown, MA, 282–289. http://dl.acm.org/citation.cfm?id=645530.655813

[21] Stephanie M. Lukin, Pranav Anand, Marilyn A. Walker, and Steve Whittaker. 2017. Argument Strength is in the Eye of the Beholder: Audience Effects in Persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Valencia, Spain, 742–753.

[22] Yi Mao and Guy Lebanon. 2007. Isotonic Conditional Random Fields and Local Sentiment Flow. In *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. C. Platt, and T. Hoffman (Eds.). MIT Press, Cambridge, MA, 961–968. http://papers.nips.cc/paper/3152-isotonic-conditional-random-fields-and-local-sentiment-flow.pdf

[23] Andreas C. Muller and Sven Behnke. 2014. PyStruct - Learning Structured Prediction in Python. *Journal of Machine Learning Research* 15 (2014), 2055–2060. http://jmlr.org/papers/v15/mueller14a.html

[24] Pradeep K. Murukannaiah, Anup K. Kalia, Pankaj R. Telang, and Munindar P. Singh. 2015. Resolving Goal Conflicts via Argumentation-Based Analysis of Competing Hypotheses. In *Proceedings of the 23rd IEEE International Requirements Engineering Conference (RE)*. IEEE Computer Society, Ottawa, 156–165. https://doi.org/10.1109/RE.2015.7320418

[25] Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying Words: Predicting Deception from Linguistic Styles. *Personality and Social Psychology Bulletin* 29, 5 (2003), 665–675. https://doi.org/10.1177/0146167203029005010 PMID: 15272998.

[26] Naoaki Okazaki. 2007. CRFsuite: A Fast Implementation of Conditional Random Fields (CRFs). http://www.chokkan.org/software/crfsuite/

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[28] James W. Pennebaker and Laura A. King. 1999. Linguistic Styles: Language Use as an Individual Difference. *Journal of Personality and Social Psychology* 77, 6 (1 12 1999), 1296–1312. https://doi.org/10.1037/0022-3514.77.6.1296

[29] Isaac Persing and Vincent Ng. 2017. Why Can't You Convince Me? Modeling Weaknesses in Unpersuasive Arguments. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. IJCAI, Melbourne, Australia, 4082–4088. https://doi.org/10.24963/ijcai.2017/570

[30] Ronaldo C. Prati, Gustavo E. A. P. A. Batista, and Maria Carolina Monard. 2011. A Survey on Graphical Methods for Classification Predictive Performance Evaluation. *IEEE Transactions on Knowledge and Data Engineering* 23, 11 (Nov 2011), 1601–1618. https://doi.org/10.1109/TKDE.2011.59

[31] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic Models for Analyzing and Detecting Biased Language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Sofia, Bulgaria). Association for Computational Linguistics, Sofia, Bulgaria, 1650–1659. http://www.aclweb.org/anthology/P13-1162

[32] Takaya Saito and Marc Rehmsmeier. 2015. The Precision-Recall Plot is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PloS One* 10, 3 (2015), e0118432.

[33] 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)*. Association for Computational Linguistics, Doha, Qatar, 46–56. https://doi.org/10.3115/v1/D14-1006

[34] 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 the 25th World Wide Web Conference (WWW)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 613–624.

[35] 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 the 27th World Wide Web Conference* (Lyon, France) *(WWW)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 945–954. https://doi.org/10.1145/3178876.3186142

[36] Serena Villata, Sahbi Benlamine, Elena Cabrio, Claude Frasson, and Fabien Gandon. 2018. Assessing Persuasion in Argumentation through Emotions and Mental States. In *Florida Artificial Intelligence Research Society Conference*. AAAI, Florida, 131–139. https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS18/paper/view/17677

[37] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 5635–5649. https://doi.org/10.18653/v1/P19-1566

[38] Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this Post Persuasive? Ranking Argumentative Comments in Online Forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Berlin, Germany). Association for Computational Linguistics, Berlin, Germany, 195–200.

https://doi.org/10.18653/v1/P16-2032

[39] Guangchao Yuan, Pradeep K. Murukannaiah, Zhe Zhang, and Munindar P. Singh. 2014. Exploiting Sentiment Homophily for Link Prediction. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys)*. ACM, Foster City, California, 17–24. https://doi.org/10.1145/2645710.2645734

[40] Zhe Zhang and Munindar P. Singh. 2014. ReNew: A Semi-Supervised Framework for Generating Domain-Specific Lexicons and Sentiment Analysis. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*. ACL, Baltimore, 542–551. https://doi.org/10.3115/v1/P14-1051