Trouble on the Horizon: Forecasting the Derailment of Online Conversations as they Develop

Jonathan P. Chang
Cornell University
jpc362@cornell.edu

Cristian Danescu-Niculescu-Mizil
Cornell University
cristian@cs.cornell.edu

Abstract

Online discussions often derail into toxic exchanges between participants. Recent efforts mostly focused on detecting antisocial behavior after the fact, by analyzing single comments in isolation. To provide more timely notice to human moderators, a system needs to preemptively detect that a conversation is heading towards derailment before it actually turns toxic. This means modeling derailment as an emerging property of a conversation rather than as an isolated utterance-level event.

Forecasting emerging conversational properties, however, poses several inherent modeling challenges. First, since conversations are dynamic, a forecasting model needs to capture the flow of the discussion, rather than properties of individual comments. Second, real conversations have an unknown horizon: they can end or derail at any time; thus a practical forecasting model needs to assess the risk in an online fashion, as the conversation develops. In this work we introduce a conversational forecasting model that learns an unsupervised representation of conversational dynamics and exploits it to predict future derailment as the conversation develops. By applying this model to two new diverse datasets of online conversations with labels for antisocial events, we show that it outperforms state-of-the-art systems at forecasting derailment.

1 Introduction

> "Ché saetta previsa vien più lenta."[1]
>
> – Dante Alighieri, Divina Commedia, Paradiso

Antisocial behavior is a persistent problem plaguing online conversation platforms; it is both widespread (Duggan, 2014) and potentially damaging to mental and emotional health (Raskauskas and Stoltz, 2007; Akbulut et al., 2010). The strain this phenomenon puts on community maintainers has sparked recent interest in computational approaches for assisting human moderators.

(1) **[User A]** What does [quote omitted] refer to? I assume it should be written from June 2010 to December 2011 and we should precise [sic] the months.
(2) **[User B]** No. It refers to 2007-2011 Belgian political crisis
(3) **[User A]** 2007-2011 Belgian political crisis is a little bit [of original research]. It merges 2 crisis in 1. Sources 2 and 4 in the article talk about a 18 months crisis in 2010-2011, ie what I refer to. What are the reliable sources that make this crisis go back to 2007?
(4) **[User B]** Yes it's not ridiculous at all to claim [it's original research] because it doesn't fit your argument. A crisis can be composed out of several smaller crisis. It's not original research if some of your sources only talk about parts [...]
(5) **[User A]** Where is the source that claim the crisis is 4 year long? Sources state claim it is 18 month long and refer to the period from June 2010 to December 2011.
(6) **[User B]** There were 4 governments and 2 years of no government in 4 years time. You can not sanely claim that this Must be viewed as two seperate crisis. What exactly splits them up? [...]

Figure 1: Example start of a conversation that will eventually derail into a personal attack.

Prior work in this direction has largely focused on *post-hoc* identification of various kinds of antisocial behavior, including hate speech (Warner and Hirschberg, 2012; Davidson et al., 2017), harassment (Yin et al., 2009), personal attacks (Wulczyn et al., 2017), and general toxicity (Pavlopoulos et al., 2017). The fact that these approaches only identify antisocial content *after the fact* limits their practicality as tools for assisting *pre-emptive* moderation in conversational domains.

Addressing this limitation requires *forecasting* the future derailment of a conversation based on early warning signs, giving the moderators time to potentially intervene before any harm is done (Liu et al. 2018, Zhang et al. 2018a, see Jurgens et al. 2019 for a discussion). Such a goal recognizes derailment as emerging from the devel-

[1] "The arrow one foresees arrives more gently."

opment of the conversation, and belongs to the broader area of *conversational forecasting*, which includes future-prediction tasks such as predicting the eventual length of a conversation (Backstrom et al., 2013), whether a persuasion attempt will eventually succeed (Tan et al., 2016; Wachsmuth et al., 2018; Yang et al., 2019), whether team discussions will eventually lead to an increase in performance (Niculae and Danescu-Niculescu-Mizil, 2016), or whether ongoing counseling conversations will eventually be perceived as helpful (Althoff et al., 2016).[2]

Approaching such *conversational forecasting* problems, however, requires overcoming several inherent modeling challenges. First, conversations are *dynamic* and their outcome might depend on how subsequent comments interact with each other. Consider the example in Figure 1: while no individual comment is outright offensive, a human reader can sense a tension emerging from their succession (e.g., dismissive answers to repeated questioning). Thus a forecasting model needs to capture not only the content of each individual comment, but also the *relations* between comments. Previous work has largely relied on hand-crafted features to capture such relations—e.g., similarity between comments (Althoff et al., 2016; Tan et al., 2016) or conversation structure (Zhang et al., 2018b; Hessel and Lee, 2019)—, though neural attention architectures have also recently shown promise (Jo et al., 2018).

The second modeling challenge stems from the fact that conversations have an *unknown horizon*: they can be of varying lengths, and the to-be-forecasted event can occur at any time. So when is it a good time to make a forecast? Prior work has largely proposed two solutions, both resulting in important practical limitations. One solution is to assume (unrealistic) prior knowledge of when the to-be-forecasted event takes place and extract features up to that point (Niculae et al., 2015; Liu et al., 2018). Another compromising solution is to extract features from a fixed-length window, often at the start of the conversation (Curhan and Pentland, 2007; Niculae and Danescu-Niculescu-Mizil, 2016; Althoff et al., 2016; Zhang et al., 2018a, inter alia). Choosing a catch-all window-size is however impractical: short windows will miss information in comments they do not encompass (e.g., a window of only two comments would miss the chain of repeated questioning in comments 3 through 6 of Figure 1), while longer windows risk missing the to-be-forecasted event altogether if it occurs before the end of the window, which would prevent *early* detection.

In this work we introduce a model for forecasting conversational events that overcomes both these inherent challenges by processing comments, and their relations, as they happen (i.e., in an online fashion). Our main insight is that models with these properties already exist, albeit geared toward generation rather than prediction: recent work in context-aware dialog generation (or "chatbots") has proposed sequential neural models that make effective use of the intra-conversational dynamics (Sordoni et al., 2015b; Serban et al., 2016, 2017), while concomitantly being able to process the conversation as it develops (see Gao et al. (2018) for a survey).

In order for these systems to perform well in the generative domain they need to be trained on massive amounts of (unlabeled) conversational data. The main difficulty in directly adapting these models to the supervised domain of conversational forecasting is the relative scarcity of labeled data: for most forecasting tasks, at most a few thousands labeled examples are available, insufficient for the notoriously data-hungry sequential neural models.

To overcome this difficulty, we propose to decouple the objective of learning a neural representation of conversational dynamics from the objective of predicting future events. The former can be *pre-trained* on large amounts of unsupervised data, similarly to how chatbots are trained. The latter can piggy-back on the resulting representation after *fine-tuning* it for classification using relatively small labeled data. While similar pre-train-then-fine-tune approaches have recently achieved state-of-the-art performance in a number of NLP tasks—including natural language inference, question answering, and commonsense reasoning (discussed in Section 2)—to the best of our knowledge this is the first attempt at applying this paradigm to conversational forecasting.

To test the effectiveness of this new architecture in forecasting derailment of online conversations, we develop and distribute two new datasets. The

[2]We can distinguish two types of forecasting tasks, depending on whether the to-be-forecasted target is an event that might take place within the conversation (e.g., derailment) or an outcome measured after the conversation will eventually conclude (e.g., helpfulness). The following discussion of modeling challenges holds for both.

first triples in size the highly curated 'Conversations Gone Awry' dataset (Zhang et al., 2018a), where civil-starting Wikipedia Talk Page conversations are crowd-labeled according to whether they eventually lead to personal attacks; the second relies on in-the-wild moderation of the popular subreddit ChangeMyView, where the aim is to forecast whether a discussion will later be subject to moderator action due to "rude or hostile" behavior. In both datasets, our model outperforms existing fixed-window approaches, as well as simpler sequential baselines that cannot account for inter-comment relations. Furthermore, by virtue of its online processing of the conversation, our system can provide substantial prior notice of upcoming derailment, triggering on average 3 comments (or 3 hours) before an overtly toxic comment is posted.

To summarize, in this work we:

- introduce the first model for forecasting conversational events that can capture the dynamics of a conversation *as it develops*;
- build two diverse datasets (one entirely new, one extending prior work) for the task of forecasting derailment of online conversations;
- compare the performance of our model against the current state-of-the-art, and evaluate its ability to provide *early* warning signs.

Our work is motivated by the goal of assisting human moderators of online communities by preemptively signaling at-risk conversations that might deserve their attention. However, we caution that any automated systems might encode or even amplify the biases existing in the training data (Park et al., 2018; Sap et al., 2019; Wiegand et al., 2019), so a public-facing implementation would need to be exhaustively scrutinized for such biases (Feldman et al., 2015).

2 Further Related Work

Antisocial behavior. Antisocial behavior online comes in many forms, including harassment (Vitak et al., 2017), cyberbullying (Singh et al., 2017), and general aggression (Kayany, 1998). Prior work has sought to understand different aspects of such behavior, including its effect on the communities where it happens (Collier and Bear, 2012; Arazy et al., 2013), the actors involved (Cheng et al., 2017; Volkova and Bell, 2017; Kumar et al., 2018; Ribeiro et al., 2018) and connections to the outside world (Olteanu et al., 2018).

Post-hoc classification of conversations. There is a rich body of prior work on classifying the outcome of a conversation after it has concluded, or classifying conversational events after they happened. Many examples exist, but some more closely related to our present work include identifying the winner of a debate (Zhang et al., 2016; Potash and Rumshisky, 2017; Wang et al., 2017), identifying successful negotiations (Curhan and Pentland, 2007; Cadilhac et al., 2013), as well as detecting whether deception (Girlea et al., 2016; Pérez-Rosas et al., 2016; Levitan et al., 2018) or disagreement (Galley et al., 2004; Abbott et al., 2011; Allen et al., 2014; Wang and Cardie, 2014; Rosenthal and McKeown, 2015) has occurred.

Our goal is different because we wish to *forecast* conversational events before they happen and while the conversation is still ongoing (potentially allowing for interventions). Note that some post-hoc tasks can also be re-framed as forecasting tasks (assuming the existence of necessary labels); for instance, predicting whether an ongoing conversation *will* eventually spark disagreement (Hessel and Lee, 2019), rather than detecting already-existing disagreement.

Conversational forecasting. As described in Section 1, prior work on forecasting conversational outcomes and events has largely relied on hand-crafted features to capture aspects of conversational dynamics. Example feature sets include statistical measures based on similarity between utterances (Althoff et al., 2016), sentiment imbalance (Niculae et al., 2015), flow of ideas (Niculae et al., 2015), increase in hostility (Liu et al., 2018), reply rate (Backstrom et al., 2013) and graph representations of conversations (Garimella et al., 2017; Zhang et al., 2018b). By contrast, we aim to automatically learn neural representations of conversational dynamics through pre-training.

Such hand-crafted features are typically extracted from fixed-length windows of the conversation, leaving unaddressed the problem of unknown horizon. While some work has trained *multiple* models for different window-lengths (Liu et al., 2018; Hessel and Lee, 2019), they consider these models to be independent and, as such, do not address the issue of aggregating them into a single forecast (i.e., deciding at what point to make

a prediction). We implement a simple sliding windows solution as a baseline (Section 5).

Pre-training for NLP. The use of pre-training for natural language tasks has been growing in popularity after recent breakthroughs demonstrating improved performance on a wide array of benchmark tasks (Peters et al., 2018; Radford et al., 2018). Existing work has generally used a language modeling objective as the pre-training objective; examples include next-word prediction (Howard and Ruder, 2018), sentence autoencoding, (Dai and Le, 2015), and machine translation (McCann et al., 2017). BERT (Devlin et al., 2019) introduces a variation on this in which the goal is to predict the next sentence in a document given the current sentence. Our pre-training objective is similar in spirit, but operates at a *conversation* level, rather than a document level. We hence view our objective as *conversational modeling* rather than (only) language modeling. Furthermore, while BERT's sentence prediction objective is framed as a multiple-choice task, our objective is framed as a generative task.

3 Derailment Datasets

We consider two datasets, representing related but slightly different forecasting tasks. The first dataset is an expanded version of the annotated Wikipedia conversations dataset from Zhang et al. (2018a). This dataset uses carefully-controlled crowdsourced labels, strictly filtered to ensure the conversations are civil up to the moment of a personal attack. This is a useful property for the purposes of model analysis, and hence we focus on this as our primary dataset. However, we are conscious of the possibility that these strict labels may not fully capture the kind of behavior that moderators care about in practice. We therefore introduce a secondary dataset, constructed from the subreddit ChangeMyView (CMV) that does not use post-hoc annotations. Instead, the prediction task is to forecast whether the conversation will be subject to moderator action in the future.

Wikipedia data. Zhang et al.'s 'Conversations Gone Awry' dataset consists of 1,270 conversations that took place between Wikipedia editors on publicly accessible talk pages. The conversations are sourced from the WikiConv dataset (Hua et al., 2018) and labeled by crowdworkers as either containing a *personal attack* from within (i.e., hostile behavior by one user in the conversation directed towards another) or remaining civil throughout.

A series of controls are implemented to prevent models from picking up on trivial correlations. To prevent models from capturing topic-specific information (e.g., political conversations are more likely to derail), each attack-containing conversation is paired with a clean conversation from the same talk page, where the talk page serves as a proxy for topic.[3] To force models to actually capture conversational dynamics rather than detecting already-existing toxicity, human annotations are used to ensure that all comments preceding a personal attack are civil.

To the ends of more effective model training, we elected to expand the 'Conversations Gone Awry' dataset, using the original annotation procedure. Since we found that the original data skewed towards shorter conversations, we focused this crowdsourcing run on longer conversations: ones with 4 or more comments preceding the attack.[4] Through this additional crowdsourcing, we expand the dataset to 4,188 conversations, which we are publicly releasing as part of the Cornell Conversational Analysis Toolkit (ConvoKit).[5]

We perform an 80-20-20 train/dev/test split, ensuring that paired conversations end up in the same split in order to preserve the topic control. Finally, we randomly sample another 1 million conversations from WikiConv to use for the unsupervised pre-training of the generative component.

Reddit CMV data. The CMV dataset is constructed from conversations collected via the Reddit API. In contrast to the Wikipedia-based dataset, we explicitly avoid the use of post-hoc annotation. Instead, we use as our label whether a conversation eventually had a comment removed by a moderator for violation of Rule 2: "Don't be rude or hostile to other users".[6]

Though the lack of post-hoc annotation limits the degree to which we can impose controls on the data (e.g., some conversations may contain toxic comments not flagged by the moderators) we do reproduce as many of the Wikipedia data's controls as we can. Namely, we replicate the topic

[3] Paired conversations were also enforced to be similar in length, so that length distribution is the same between classes.

[4] We cap the length at 10 to avoid overwhelming the crowdworkers.

[5] convokit.cornell.edu

[6] The existence of this specific rule, the standardized moderation messages and the civil character of the ChangeMyView subreddit was our initial motivation for choosing it.

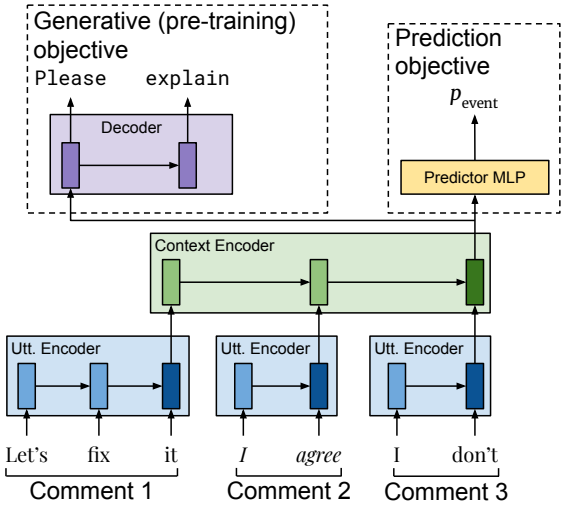

Figure 2: Sketch of the CRAFT architecture.

control pairing by choosing pairs of positive and negative examples that belong to the same top-level post, following Tan et al. (2016);[7] and enforce that the removed comment was made by a user who was previously involved in the conversation.[8] This process results in 6,842 conversations, to which we again apply a pair-preserving 80-20-20 split. Finally, we gather over 600,000 conversations that do not include any removed comment, for unsupervised pre-training.

4 Conversational Forecasting Model

We now describe our general model for forecasting future conversational events. Our model integrates two components: (a) a generative dialog model that learns to represent conversational dynamics in an unsupervised fashion; and (b) a supervised component that fine-tunes this representation to forecast future events. Figure 2 provides an overview of the proposed architecture, henceforth CRAFT (Conversational Recurrent Architecture for ForecasTing).

Terminology. For modeling purposes, we treat a conversation as a sequence of N comments $C = \{c_1, \ldots, c_N\}$. Each comment, in turn, is a sequence of tokens, where the number of tokens may vary from comment to comment. For the n-th comment ($1 \leq n \leq N$), we let M_n denote the number of tokens. Then, a comment c_n can be represented as a sequence of M_n tokens: $c_n = \{w_1, \ldots, w_{M_n}\}$.

[7] The top-level post is not part of the conversations.

[8] We also impose the same length restriction on the number of comments preceding the removed comment, for comparability and for computational considerations.

Generative component. For the generative component of our model, we use a hierarchical recurrent encoder-decoder (HRED) architecture (Sordoni et al., 2015a), a modified version of the popular sequence-to-sequence (seq2seq) architecture (Sutskever et al., 2014) designed to account for dependencies between consecutive inputs. Serban et al. (2016) showed that HRED can successfully model conversational context by encoding the temporal structure of previously seen comments, making it an ideal fit for our use case. Here, we provide a high-level summary of the HRED architecture, deferring deeper technical discussion to Sordoni et al. (2015a) and Serban et al. (2016).

An HRED dialog model consists of three components: an utterance encoder, a context encoder, and a decoder. The utterance encoder is responsible for generating semantic vector representations of comments. It consists of a recurrent neural network (RNN) that reads a comment token-by-token, and on each token w_m updates a hidden state h^{enc} based on the current token and the previous hidden state:

$$h_m^{\text{enc}} = f^{\text{RNN}}(h_{m-1}^{\text{enc}}, w_m) \tag{1}$$

where f^{RNN} is a nonlinear gating function (our implementation uses GRU (Cho et al., 2014)). The final hidden state h_M^{enc} can be viewed as a vector encoding of the entire comment.

Running the encoder on each comment c_n results in a sequence of N vector encodings. A second encoder, the context encoder, is then run over this sequence:

$$h_n^{\text{con}} = f^{\text{RNN}}(h_{n-1}^{\text{con}}, h_{M_n}^{\text{enc}}) \tag{2}$$

Each hidden state h_n^{con} can then be viewed as an encoding of the full conversational context up to and including the n-th comment. To generate a response to comment n, the context encoding h_n^{con} is used to initialize the hidden state h_0^{dec} of a decoder RNN. The decoder produces a response token by token using the following recurrence:

$$\begin{aligned} h_t^{\text{dec}} &= f^{\text{RNN}}(h_{t-1}^{\text{dec}}, w_{t-1}) \\ w_t &= f^{\text{out}}(h_t^{\text{dec}}) \end{aligned} \tag{3}$$

where f^{out} is some function that outputs a probability distribution over words; we implement this using a simple feedforward layer. In our implementation, we further augment the decoder with attention (Bahdanau et al., 2014; Luong et al.,

2015) over context encoder states to help capture long-term inter-comment dependencies. This generative component can be pre-trained using unlabeled conversational data.

Prediction component. Given a pre-trained HRED dialog model, we aim to extend the model to predict from the conversational context whether the to-be-forecasted event will occur. Our predictor consists of a multilayer perceptron (MLP) with 3 fully-connected layers, leaky ReLU activations between layers, and sigmoid activation for output. For each comment c_n, the predictor takes as input the context encoding h_n^{con} and forwards it through the MLP layers, resulting in an output score that is interpreted as a probability $p_{\text{event}}(c_{n+1})$ that the to-be-forecasted event will happen (e.g., that the conversation will derail).

Training the predictive component starts by initializing the weights of the encoders to the values learned in pre-training. The main training loop then works as follows: for each positive sample—i.e., a conversation containing an instance of the to-be-forecasted event (e.g., derailment) at comment c_e—we feed the context $c_1, \ldots, c_{e-1}$ through the encoder and classifier, and compute cross-entropy loss between the classifier output and expected output of 1. Similarly, for each negative sample—i.e., a conversation where none of the comments exhibit the to-be-forecasted event and that ends with c_N—we feed the context $c_1, \ldots, c_{N-1}$ through the model and compute loss against an expected output of 0.

Note that the parameters of the generative component are not held fixed during this process; instead, backpropagation is allowed to go all the way through the encoder layers. This process, known as *fine-tuning*, reshapes the representation learned during pre-training to be more directly useful to prediction (Howard and Ruder, 2018).

We implement the model and training code using PyTorch, and we are publicly releasing our implementation and the trained models together with the data as part of ConvoKit.

5 Forecasting Derailment

We evaluate the performance of CRAFT in the task of forecasting conversational derailment in both the Wikipedia and CMV scenarios. To this end, for each of these datasets we pre-train the generative component on the unlabeled portion of the data and fine-tune it on the labeled training split (data size detailed in Section 3).

In order to evaluate our sequential system against conversational-level ground truth, we need to aggregate comment level predictions. If *any* comment in the conversation *triggers* a positive prediction—i.e., $p_{\text{event}}(c_{n+1})$ is greater than a threshold learned on the development split—then the respective conversation is predicted to derail. If this forecast is triggered in a conversation that actually derails, but *before* the derailment actually happens, then the conversation is counted as a true positive; otherwise it is a false positive. If no positive predictions are triggered for a conversation, but it actually derails then it counts as a false negative; if it does not derail then it is a true negative.

Fixed-length window baselines. We first seek to compare CRAFT to existing, fixed-length window approaches to forecasting. To this end, we implement two such baselines: *Awry*, which is the state-of-the-art method proposed in Zhang et al. (2018a) based on pragmatic features in the first comment-reply pair,[9] and *BoW*, a simple bag-of-words baseline that makes a prediction using TF-IDF weighted bag-of-words features extracted from the first comment-reply pair.

Online forecasting baselines. Next, we consider simpler approaches for making forecasts as the conversations happen (i.e., in an online fashion). First, we propose *Cumulative BoW*, a model that recomputes bag-of-words features on all comments seen thus far every time a new comment arrives. While this approach does exhibit the desired behavior of producing updated predictions for each new comment, it fails to account for relationships between comments.

This simple cumulative approach cannot be directly extended to models whose features are strictly based on a fixed number of comments, like Awry. An alternative is to use a *sliding window*: for a feature set based on a window of W comments, upon each new comment we can extract features from a window containing that comment and the $W-1$ comments preceding it. We apply this to the Awry method and call this model *Sliding Awry*. For both these baselines, we aggregate comment-level predictions in the same way as in our main model.

CRAFT ablations. Finally, we consider two modified versions of the CRAFT model in order

[9] We use the ConvoKit implementation.

	Capabilities			Wikipedia Talk Pages					Reddit CMV				
Model	**D**	**O**	**L**	**A**	**P**	**R**	**FPR**	**F1**	**A**	**P**	**R**	**FPR**	**F1**
BoW				56.5	55.6	65.5	52.4	60.1	52.1	51.8	61.3	57.0	56.1
Awry	✓			58.9	59.2	57.6	39.8	58.4	54.4	55.0	48.3	**39.5**	51.4
Cumul. BoW		✓		60.6	57.7	**79.3**	58.1	66.8	59.9	58.8	65.9	46.2	62.1
Sliding Awry	✓	✓		60.6	60.2	62.4	41.2	61.3	56.8	56.6	58.2	44.6	57.4
CRAFT − CE		✓	✓	64.9	**64.4**	66.7	**36.9**	65.5	57.7	56.1	71.2	55.7	62.8
CRAFT	✓	✓	✓	**66.5**	63.7	77.1	44.1	**69.8**	**63.4**	**60.4**	**77.5**	50.7	**67.9**

Table 1: Comparison of the capabilities of each baseline and our CRAFT models (full and without the Context Encoder) with regards to capturing inter-comment (D)ynamics, processing conversations in an (O)nline fashion, and automatically (L)earning feature representations, as well as their performance in terms of (A)ccuracy, (P)recision, (R)ecall, False Positive Rate (FPR), and F1 score. Awry is the model previously proposed by Zhang et al. (2018a) for this task.

to evaluate the impact of two of its key components: (1) the pre-training step, and (2) its ability to capture inter-comment dependencies through its hierarchical memory.

To evaluate the impact of pre-training, we train the prediction component of CRAFT on only the labeled training data, without first pre-training the encoder layers with the unlabeled data. We find that given the relatively small size of labeled data, this baseline fails to successfully learn, and ends up performing at the level of random guessing.[10] This result underscores the need for the pre-training step that can make use of unlabeled data.

To evaluate the impact of the hierarchical memory, we implement a simplified version of CRAFT where the memory size of the context encoder is zero (*CRAFT − CE*), thus effectively acting as if the pre-training component is a vanilla seq2seq model. In other words, this model cannot capture inter-comment dependencies, and instead at each step makes a prediction based only on the utterance encoding of the latest comment.

Results. Table 1 compares CRAFT to the baselines on the test splits (random baseline is 50%) and illustrates several key findings. First, we find that unsurprisingly, accounting for full conversational context is indeed helpful, with even the simple online baselines outperforming the fixed-window baselines. On both datasets, CRAFT outperforms all baselines (including the other online models) in terms of accuracy and F1. Furthermore, although it loses on precision (to CRAFT − CE) and recall (to Cumulative BoW) individually on the Wikipedia data, CRAFT has the superior *balance* between the two, having both a visibly higher precision-recall curve and larger area under the curve (AUPR) than the baselines (Figure 3). This latter property is particularly useful in a practical setting, as it allows moderators to tune model performance to some desired precision without having to sacrifice as much in the way of recall (or vice versa) compared to the baselines and pre-existing solutions.

[10] We thus exclude this baseline from the results summary.

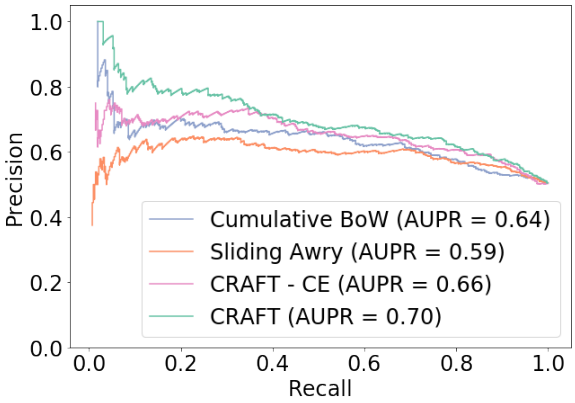

Figure 3: Precision-recall curves and the area under each curve. To reduce clutter, we show only the curves for Wikipedia data (CMV curves are similar) and exclude the fixed-length window baselines (which perform worse).

6 Analysis

We now examine the behavior of CRAFT in greater detail, to better understand its benefits and limitations. We specifically address the following questions: (1) How much early warning does the the model provide? (2) Does the model actually

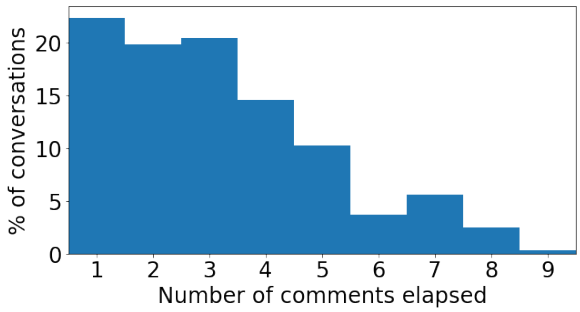

Figure 4: Distribution of number of comments elapsed between the model's first warning and the attack.

learn an order-sensitive representation of conversational context?[11]

Early warning, but how early? The recent interest in forecasting antisocial behavior has been driven by a desire to provide pre-emptive, actionable warning to moderators. But does our model trigger early enough for any such practical goals?

For each personal attack correctly forecasted by our model, we count the number of comments elapsed between the time the model is first triggered and the attack. Figure 4 shows the distribution of these counts: on average, the model warns of an attack 3 comments before it actually happens (4 comments for CMV). To further evaluate how much time this early warning would give to the moderator, we also consider the difference in timestamps between the comment where the model first triggers and the comment containing the actual attack. Over 50% of conversations get at least 3 hours of advance warning (2 hours for CMV). Moreover, 39% of conversations get at least *12 hours* of early warning before they derail.

Does order matter? One motivation behind the design of our model was the intuition that comments in a conversation are not independent events; rather, the order in which they appear matters (e.g., a blunt comment followed by a polite one feels intuitively different from a polite comment followed by a blunt one). By design, CRAFT has the capacity to learn an order-sensitive representation of conversational context, but how can we know that this capacity is actually used? It is conceivable that the model is simply computing an order-insensitive "bag-of-features". Neural network models are notorious for their lack of transparency, precluding an analysis of how *exactly* CRAFT models conversational context. Nevertheless, through two simple exploratory experiments, we seek to show that it does not completely ignore comment order.

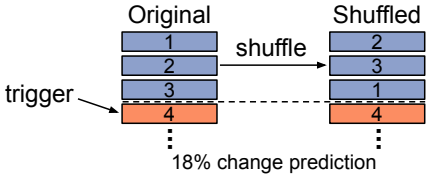

Figure 5: The prefix-shuffling procedure ($t = 4$).

The first experiment for testing whether the model accounts for comment order is a *prefix-shuffling* experiment, visualized in Figure 5. For each conversation that the model predicts will derail, let t denote the index of the triggering comment, i.e., the index where the model first made a derailment forecast. We then construct *synthetic* conversations by taking the first $t-1$ comments (henceforth referred to as the *prefix*) and randomizing their order.[12] Finally, we count how often the model no longer predicts derailment at index t in the synthetic conversations. If the model were ignoring comment order, its prediction should remain unchanged (as it remains for the Cumulative BoW baseline), since the actual *content* of the first t comments has not changed (and CRAFT inference is deterministic). We instead find that in roughly one fifth of cases (12% for CMV) the model changes its prediction on the synthetic conversations. This suggests that CRAFT learns an order-sensitive representation of context, not a mere "bag-of-features".

To more concretely quantify how much this order-sensitive context modeling helps with prediction, we can actively prevent the model from learning and exploiting any order-related dynamics. We achieve this through another type of shuffling experiment, where we go back even further and shuffle the comment order in the conversations used for pre-training, fine-tuning and testing. This procedure preserves the model's ability to capture signals present within the individual comments processed so far, as the utterance encoder is unaffected, but inhibits it from capturing any meaningful order-sensitive dynamics. We find that this hurts the model's performance (65% ac-

[11] We choose to focus on the Wikipedia scenario since the conversational prefixes are hand-verified to be civil. For completeness we also report results for Reddit CMV throughout, but they should be taken with an additional grain of salt.

[12] We restrict the experiment to cases where $t \geq 3$, as prefixes consisting of only one comment cannot be reordered.

curacy for Wikipedia, 59.5% for CMV), lowering it to a level similar to that of the version where we completely disable the context encoder.

Taken together, these experiments provide evidence that CRAFT uses its capacity to model conversational context in an order-sensitive fashion, and that it makes effective use of the dynamics within. An important avenue for future work would be developing more transparent models that can shed light on exactly *what* kinds of order-related features are being extracted and *how* they are used in prediction.

7 Conclusions and Future Work

In this work, we introduced a model for forecasting conversational events that processes comments as they happen and takes the full conversational context into account to make an updated prediction at each step. This model fills a void in the existing literature on conversational forecasting, simultaneously addressing the dual challenges of capturing inter-comment dynamics and dealing with an unknown horizon. We find that our model achieves state-of-the-art performance on the task of forecasting derailment in two different datasets that we release publicly. We further show that the resulting system can provide substantial prior notice of derailment, opening up the potential for preemptive interventions by human moderators (Seering et al., 2017).

While we have focused specifically on the task of forecasting derailment, we view this work as a step towards a more general model for real-time forecasting of other types of emergent properties of conversations. Follow-up work could adapt the CRAFT architecture to address other forecasting tasks mentioned in Section 2—including those for which the outcome is extraneous to the conversation. We expect different tasks to be informed by different types of inter-comment dynamics, and further architecture extensions could add additional supervised fine-tuning in order to direct it to focus on specific dynamics that might be relevant to the task (e.g., exchange of ideas between interlocutors or stonewalling).

With respect to forecasting derailment, there remain open questions regarding what human moderators actually desire from an early-warning system, which would affect the design of a practical system based on this work. For instance, how early does a warning need to be in order for moderators to find it useful? What is the optimal balance between precision, recall, and false positive rate at which such a system is truly improving moderator productivity rather than wasting their time through false positives? What are the ethical implications of such a system? Follow-up work could run a user study of a prototype system with actual moderators to address these questions.

A practical limitation of the current analysis is that it relies on balanced datasets, while derailment is a relatively rare event for which a more restrictive trigger threshold would be appropriate. While our analysis of the precision-recall curve suggests the system is robust across multiple thresholds ($AUPR = 0.7$), additional work is needed to establish whether the recall tradeoff would be acceptable in practice.

Finally, one major limitation of the present work is that it assigns a single label to each conversation: does it derail or not? In reality, derailment need not spell the end of a conversation; it is possible that a conversation could get back on track, suffer a repeat occurrence of antisocial behavior, or any number of other trajectories. It would be exciting to consider finer-grained forecasting of conversational trajectories, accounting for the natural—and sometimes chaotic—ebb-and-flow of human interactions.

Acknowledgements. We thank Caleb Chiam, Liye Fu, Lillian Lee, Alexandru Niculescu-Mizil, Andrew Wang and Justine Zhang for insightful conversations (with unknown horizon), Aditya Jha for his great help with implementing and running the crowd-sourcing tasks, Thomas Davidson and Claire Liang for exploratory data annotation, as well as the anonymous reviewers for their helpful comments. This work is supported in part by the NSF CAREER award IIS-1750615 and by the NSF Grant SES-1741441.

References

Rob Abbott, Marilyn Walker, Pranav Anand, Jean E. Fox Tree, Robeson Bowmani, and Joseph King. 2011. How Can You Say Such Things?!?: Recognizing Disagreement in Informal Political Argument. In *Proceedings of the Workshop on Languages in Social Media*.

Yavuz Akbulut, Yusuf Levent Sahin, and Bahadir Eristi. 2010. Cyberbullying Victimization among Turkish Online Social Utility Members. *Educational Technology & Society*, 13(4).

Kelsey Allen, Giuseppe Carenini, and Raymond T. Ng. 2014. Detecting Disagreement in Conversations using Pseudo-Monologic Rhetorical Structure. In *Proceedings of EMNLP*.

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics*, 4.

Ofer Arazy, Lisa Yeo, and Oded Nov. 2013. Stay on the Wikipedia Task: When Task-related Disagreements Slip Into Personal and Procedural Conflicts. *J. Am. Soc. Inf. Sci. Technol.*, 64(8).

Lars Backstrom, Jon Kleinberg, Lillian Lee, and Cristian Danescu-Niculescu-Mizil. 2013. Characterizing and Curating Conversation Threads: Expansion, Focus, Volume, Re-entry. In *Proceedings of WSDM*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of ICLR*.

Anais Cadilhac, Nicholas Asher, Farah Benamara, and Alex Lascarides. 2013. Grounding Strategic Conversation: Using Negotiation Dialogues to Predict Trades in a Win-Lose Game. In *Proceedings of EMNLP*.

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of CSCW*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of EMNLP*.

Benjamin Collier and Julia Bear. 2012. Conflict, Criticism, or Confidence: An Empirical Examination of the Gender Gap in Wikipedia Contributions. In *Proceedings of CSCW*.

Jared R. Curhan and Alex Pentland. 2007. Thin Slices of Negotiation: Predicting Outcomes From Conversational Dynamics Within the First 5 Minutes. *Journal of Applied Psychology*, 92.

Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised Sequence Learning. In *Proceedings of NeurIPS*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of ICWSM*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Maeve Duggan. 2014. Online Harassment. http://www.pewinternet.org/2014/10/22/online-harassment/.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of KDD*.

Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. 2004. Identifying Agreement and Disagreement in Conversational Speech: Use of Bayesian Networks to Model Pragmatic Dependencies. In *Proceedings of ACL*.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural Approaches to Conversational AI. In *Proceedings of SIGIR*.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2017. Quantifying Controversy in Social Media. *ACM Transactions on Social Computing*, 1(1).

Codruta Girlea, Roxana Girju, and Eyal Amir. 2016. Psycholinguistic Features for Deceptive Role Detection in Werewolf. In *Proceedings of NAACL*.

Jack Hessel and Lillian Lee. 2019. Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features. In *Proceedings of NAACL*.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of ACL*.

Yiqing Hua, Cristian Danescu-Niculescu-Mizil, Dario Taraborelli, Nithum Thain, Jeffery Sorensen, and Lucas Dixon. 2018. WikiConv: A Corpus of the Complete Conversational History of a Large Online Collaborative Community. In *Proceedings of EMNLP*.

Yohan Jo, Shivani Poddar, Byungsoo Jeon, Qinlan Shen, Carolyn P. Rosé, and Graham Neubig. 2018. Attentive Interaction Model: Modeling Changes in View in Argumentation. In *Proceedings of NAACL*.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. In *Proceedings of ACL*.

Joseph M. Kayany. 1998. Contexts of uninhibited online behavior: Flaming in social newsgroups on usenet. *Journal of the American Society for Information Science*, 49(12).

Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community Interaction and Conflict on the Web. In *Proceedings of WWW*.

Sarah Ita Levitan, Angel Maredia, and Julia Hirschberg. 2018. Linguistic Cues to Deception and Perceived Deception in Interview Dialogues. In *Proceedings of NAACL*.

Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. 2018. Forecasting the Presence and Intensity of Hostility on Instagram Using Linguistic and Social Features. In *Proceedings of ICWSM*.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of EMNLP*.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in Translation: Contextualized Word Vectors. In *Proceedings of NeurIPS*.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2016. Conversational Markers of Constructive Discussions. In *Proceedings of NAACL*.

Vlad Niculae, Srijan Kumar, Jordan Boyd-Graber, and Cristian Danescu-Niculescu-Mizil. 2015. Linguistic Harbingers of Betrayal: A Case Study on an Online Strategy Game. In *Proceedings of ACL*.

Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush Varshney. 2018. The Effect of Extremist Violence on Hateful Speech Online. In *Proceedings of ICWSM*.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of EMNLP*.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper Attention to Abusive User Content Moderation. In *Proceedings of EMNLP*.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, Yao Xiao, C. J. Linton, and Mihai Burzo. 2016. Verbal and Nonverbal Clues for Real-life Deception Detection. In *Proceedings of EMNLP*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of NAACL*.

Peter Potash and Anna Rumshisky. 2017. Towards Debate Automation: A Recurrent Model for Predicting Debate Winners. In *Proceedings of EMNLP*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-training. Technical report, OpenAI.

Juliana Raskauskas and Ann D. Stoltz. 2007. Involvement in Traditional and Electronic Bullying Among Adolescents. *Developmental Psychology*, 43(3).

Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, Virgílio A. F. Almeida, and Wagner Meira Jr. 2018. Characterizing and Detecting Hateful Users on Twitter. In *Proceedings of ICWSM*.

Sara Rosenthal and Kathleen McKeown. 2015. I Couldn't Agree More: The Role of Conversational Structure in Agreement and Disagreement Detection in Online Discussions. In *Proceedings of SIGDIAL*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of ACL*.

Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of CSCW*.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *Proceedings of AAAI*.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *Proceedings of AAAI*.

Vivek K. Singh, Marie L. Radford, Qianjia Huang, and Susan Furrer. 2017. "They basically like destroyed the school one day": On Newer App Features and Cyberbullying in Schools. In *Proceedings of CSCW*.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015a. A Hierarchical Recurrent Encoder-Decoder for Generative Context-Aware Query Suggestion. In *Proceedings of CIKM*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015b. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. In *Proceedings of NAACL*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of NeurIPS*.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu, and Lillian Lee. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *Proceedings of WWW*.

Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying Women's Experiences With and Strategies for Mitigating Negative Effects of Online Harassment. In *Proceedings of CSCW*.

Svitlana Volkova and Eric Bell. 2017. Identifying Effective Signals to Predict Deleted and Suspended Accounts on Twitter across Languages. In *Proceedings of ICWSM*.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the Best Counterargument without Prior Topic Knowledge. In *Proceedings of ACL*.

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.

Lu Wang and Claire Cardie. 2014. A Piece of My Mind: A Sentiment Analysis Approach for Online Dispute Detection. In *Proceedings of ACL*.

William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: The Problem of Biased Datasets. In *Proceedings of NAACL*.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of WWW*.

Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. 2019. Let's Make Your Request More Persuasive: Modeling Persuasive Strategies via Semi-Supervised Neural Nets on Crowdfunding Platforms. In *Proceedings of NAACL*.

Dawei Yin, Zhenzhen Xue, and Liangjie Hong. 2009. Detection of Harassment on Web 2.0. In *Proceedings of CAW2.0*.

Justine Zhang, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Nithum Thain, Yiqing Hua, and Dario Taraborelli. 2018a. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of ACL*.

Justine Zhang, Cristian Danescu-Niculescu-Mizil, Christina Sauper, and Sean J. Taylor. 2018b. Characterizing Online Public Discussions Through Patterns of Participant Interactions. In *Proceedings of CSCW*.

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