Conversation Modeling to Predict Derailment

Jiaqing Yuan and **Munindar P. Singh**

North Carolina State University
Raleigh, North Carolina
jyuan23@ncsu.edu, mpsingh@ncsu.edu

Abstract

Conversations among online users sometimes *derail*, i.e., break down into personal attacks. Derailment interferes with the healthy growth of communities in cyberspace. The ability to predict whether an ongoing conversation will derail could provide valuable advance, even real-time, insight to both interlocutors and moderators. Prior approaches predict conversation derailment retrospectively without the ability to forestall the derailment proactively. Some existing works attempt to make dynamic predictions as the conversation develops, but fail to incorporate multisource information, such as conversational structure and distance to derailment.

We propose a hierarchical transformer-based framework that combines utterance-level and conversation-level information to capture fine-grained contextual semantics. We propose a domain-adaptive pretraining objective to unite conversational structure information and a multitask learning scheme to leverage the distance from each utterance to derailment. An evaluation of our framework on two conversation derailment datasets shows an improvement in F1 score for the prediction of derailment. These results demonstrate the effectiveness of incorporating multisource information for predicting the derailment of a conversation.

1 Introduction

Online social platforms provide great opportunities for users to constructively converse and collaboratively develop ideas (Hua et al. 2018; Guo, Zhang, and Singh 2020). However, antisocial behaviors such as personal attacks impede the building of healthy and thriving online communities (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015).

Recently, Natural Language Processing (NLP) has been applied to address antisocial behavior. However, most previous research aims only at detecting antisocial behavior once the misconduct has occurred (Chandrasekharan et al. 2017; Kumar, Cheng, and Leskovec 2017). Such post hoc identification limits the actions platform moderators can take. What most platform moderators do after the detection of antisocial behavior is to remove uncivil content or suspend the poster's account. However, the damage is already done by that time. And, the involved parties may feel discouraged from participating in future conversations. Another practical problem is that uncivil content could potentially be overlooked by the moderator.

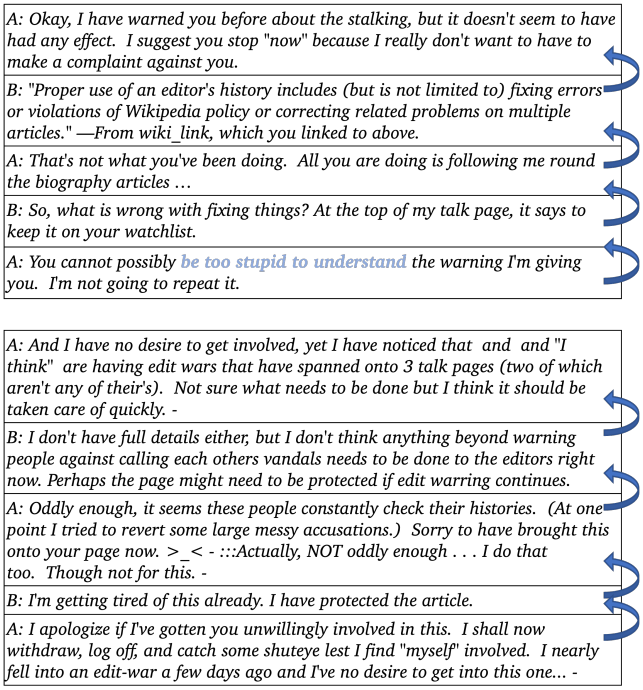

Figure 1: A pair of conversations with one developing into a personal attack in the end (top) and one staying civil throughout (bottom). Arrows show the "reply_to" relationship between utterances.

A more beneficial strategy would be to apply NLP to provide an early warning, possibly to interfere with the potential derailment of the conversation, as the conversation is developing. Figure 1 shows a pair of example conversations. Neither has personal attacks at the beginning, but one ends up with a personal attack (top) and another stays civil throughout (bottom). As an example intervention, the moderator may advocate for politeness or emphasize the rules of the platform when a (potential) derailment of an ongoing conversation is predicted. We say "potential" because derailment may be avoided. Achieving this strategy requires the model to learn and predict the dynamics of developing con-

versations, as opposed to the post hoc classification of past conversations.

Forecasting the trajectory of a conversation could enable use cases other than of detecting derailment. For example, we might adapt these methods to predict whether a conversation that persuades people to donate to charity will likely succeed or not (Wang et al. 2019), or predict whether a conversation is likely to change another person's view (Guo, Zhang, and Singh 2020).

Conversation modeling and forecasting expose important challenges. These challenges are illustrated by the two datasets we investigate. These two datasets are real conversations extracted from Wikipedia talk page and Reddit sbureddit ChangeMyView.

Dynamics There are complex dynamics at the levels of both the utterance and the conversation. The semantics of the conversation is affected not only by the content of the utterances but also by the tree-like structure between utterances. The dynamics could change abruptly because of new utterances.

Length The number of utterances that will occur in a conversation is unknown. The conversation could stop at any time and a personal attack could happen at any moment as the conversation unfolds. An earlier warning is obviously better than a warning that comes up right before the attack but earlier warnings would have lower accuracy. When should the model make a prediction and how to trade off between timeliness aspect and accuracy?

Complexity Conversation modeling challenges deep learning models because all utterances in a conversation need to be processed. Hence, the total length of a tokenized conversation produced by concatenating all utterances can be much longer than a single utterance, and can exceed the maximum input length limit (512) for BERT (Devlin et al. 2019). For instance, the average tokenized lengths for the two datasets we work with are 633 and 823, respectively.

Previous work addresses these challenges with different strategies. For example, Hessel and Lee (2019) mostly rely on hand-crafted features to model a conversation. Chang and Danescu-Niculescu-Mizil (2019) apply the LSTM architecture to capture the dynamics and consider only the first 80 tokens of each utterance. Moreover, previous works (Chang and Danescu-Niculescu-Mizil 2019; Janiszewski, Lango, and Stefanowski 2021) solely rely on textual semantics and disregard information such as conversational structure.

To investigate conversation modeling for derailment prediction, we raise three interrelated research questions:

RQ1 Is it effective to leverage pretrained language models for conversation modeling tasks and in what way?

RQ2 How can we leverage the information inherent in a conversation, such as distance from each utterance to the derailing utterance to enhance the prediction?

RQ3 Does conversation structure matter for the derailment prediction and how do we integrate it into the model?

We answer these questions with a hierarchical transformer framework that tackles the above challenges by leveraging pretrained language model. We design ways to integrate various sources of information and explore how each component other than textual content contributes to the modeling of an ongoing conversation. Specifically, we propose a multitask training scheme to leverages the time for a conversation to derail as a factor, and a pretraining scheme to use conversational structure.

2 Related Work

2.1 Antisocial Behavior in Cyberspace

Previous research defines and detects various aspects of antisocial behaviors in online platforms. These antisocial behaviors, include toxicity (Pavlopoulos et al. 2020; Ive, Anuchitanukul, and Specia 2021), abusive language and content (Vidgen et al. 2019), hate speech (Fortuna and Nunes 2018; Mozafari, Farahbakhsh, and Crespi 2019), trolling (Mojica 2017), offense (Meaney et al. 2021), and racism (Field et al. 2021). Earlier work primarily relies on hand-crafted features (Hessel and Lee 2019; Zhang et al. 2018), whereas recent work takes advantage of deep neural networks (Chang and Danescu-Niculescu-Mizil 2019). In contrast to the detection of antisocial behaviors, another line of work (Bao et al. 2021) takes a different perspective to study early cues and design metrics for quantifying and predicting prosocial outcomes in online conversations.

Much of the past work focuses on classifying the type of antisocial behavior retrospectively with a single piece of text without considering the context. Zhang et al. (2018) propose a task to predict whether an ongoing conversation will evolve into a personal attack as it develops. Pavlopoulos et al. (2020) detect and measure toxicity in context by considering the parent of a comment. Our work is a natural extension of previous work focusing on exploiting inherent contextual information to make fine-grained predictions of the trajectory of an ongoing conversation.

2.2 Conversation Modeling

Dialogue modeling is a promising line of research. Khanpour, Guntakandla, and Nielsen (2016) address the classification of dialogue acts, which involves giving a predefined act type to each utterance. However, this type of classification focuses on utterance-level prediction. Sordoni et al. (2015) propose a widely used architecture for conversation modeling that applies a hierarchical recurrent neural network for encoding an utterance and its context, respectively. Chang and Danescu-Niculescu-Mizil (2019) leverage the same architecture and pretrain the model with domain data comprising over 1 million conversations. (Guo and Singh 2023) propose a general framework for modelling argumentative relevance in conversations.

Our work differs from previous research in two aspects. First, we explore new ways of modeling conversation data by leveraging pretrained languages models. Second, we propose a new pretraining goal to incorporate the inherent tree structure of conversations into the model and evaluate its effectiveness.

2.3 Domain Adaptive Pretraining

Since the advent of BERT (Devlin et al. 2019), the pretraining-then-finetuning paradigm has been popular (Howard and Ruder 2018). However, this paradigm does not work well on some domain-specific tasks due to the lack of annotated data. Gururangan et al. (2020) propose another advanced computing paradigm, pretraining, domain-adaptive pretraining, and finetuning, to leverage the unlabeled domain-specific data. During domain adaptive pretraining, the training scheme is usually the same as during general pretraining, that is, using a masked language model.

We follow the line of thinking illustrated by Gururangan et al. (2020) but pretrain the model to identify the parent comment each utterance replies to.

3 Methodology

We now describe our model for evaluating and forecasting conversation development by integrating multisource information. We experiment with four settings, where additional information is consolidated incrementally during training to resolve each of our research questions:

BERT Kementchedjhieva and Søgaard (2021) apply a simple BERT model on the same datasets as us. They concatenate all utterances and add a classification head on top of BERT. We follow the descriptions in their work to conduct the same experiment.

Hierarchical-Base For **RQ1**, we leverage the pretrained language model to design a hierarchical transformer model that encodes the utterance-level and the conversation-level information, respectively.

Hierarchical-Multi For **RQ2**, on top of Hierarchical-Base, we propose a multitask learning scheme and leverage the distance from each utterance to the derailing utterance as an auxiliary training objective. An intuition is that the distance till derailment can provide a fine-grained signal to the model.

Hierarchical-Multi+Pretrain For **RQ3**, we take advantage of the inherent utterance structure, as captured by the "reply-to" attribute for each utterance. Each conversation is viewed as a tree structure. We set up a pretraining objective to predict the parent of each utterance.

3.1 Problem Formulation

We now define the problem formally. A conversation is a sequence of utterances, $C = \{u_1, \ldots, u_n\}$, where n is the number of utterances in the conversation. Each utterance consists of a sequence of words, $u = \{w_1, \ldots, w_m\}$, where m is the number of words in the utterance. Each conversation comes with a label $d =$ positive or $d =$ negative, where positive denotes there is a personal attack (derailment) at utterance u_n, and negative denotes the conversation is civil throughout. A data sample can be represented as a tuple $(\{u_1, \ldots, u_n\}, d)$.

We focus on predicting the possibilities of derailment for *ongoing* and *civil* conversations, i.e., how likely a civil conversation is to lead to a personal attack as it develops. To this end, exchanges between speakers after the first derailment are ignored when preparing the dataset so that the model is fed with civil utterances during training and inference. Therefore, positive samples are in the format of $\{0, \ldots, 0, 1\}$, and negative samples are in the format of $\{0, \ldots, 0, 0\}$, whereas 1 represents personal attack utterance and 0 represents civil utterance.

We don't consider what happens after a personal attack, and defer it to future research. The evaluation process is dynamic, which means that for each conversation, the model makes sequential predictions for a list of inputs, comprising increasing sequences: $(\{u_1\}, \{u_1, u_2\}, \ldots, \{u_1, u_2, \ldots, u_{n-1}\})$. The model stops whenever a positive prediction is made, which indicates that the conversation will derail. Naturally, for a conversation with label negative, the model should make $n - 1$ negative predictions.

The four model settings can be categorized into two types. The first one is a plain BERT model that concatenates all utterances and doesn't differentiate between utterance-level and conversation-level encoding. The other three models are variants of a hierarchical transformer. Each of these models encodes each utterance first, and then applies attention layers to capture conversation-level information. Below is a detailed description for the architecture.

3.2 Utterance-Level Encoder

For each utterance $u = \{w_1, \ldots, w_m\}$, we leverage the pretrained language model to capture rich semantics. Specifically, we use a transformer-based model with the same configuration as RoBERTa-base (Liu et al. 2019), initialized with pretrained weights from Huggingface.[1] RoBERTa improves over BERT by employing dynamic masking with ten times as much training data. We follow the preprocessing steps to tokenize the utterance and append special tokens [CLS] at the front and [SEP] at the end. Before feeding the token embedding into the first-layer transformer, we add a pretrained positional embedding to each token. The maximum input length for RoBERTa is 512 and we cut off extra tokens if the tokenized utterance length exceeds the limit. Finally, we take the embedding of the special token [CLS] from the last layer's output as utterance representation.

3.3 Conversation-Level Encoder

Derailment should not be considered a singular attribute of an utterance as it is the result of an entire conversation. Therefore, we consider the cumulative effect of previous utterances. For each conversation $C = \{u_1, \ldots, u_n\}$, we obtain the utterance embedding E from the first-level transformer, and then feed the sequence of utterance embeddings to four identical transformer layers. Similarly, we use the [CLS] embedding from the last layer as a representation of the entire conversation and feed it to a classifier. The classifier is made up of one fully connected linear layer for the binary classification head. We choose transformer layers over an LSTM layer because multihead attention mechanisms have an edge over the traditional LSTM model. To

[1] https://huggingface.co/

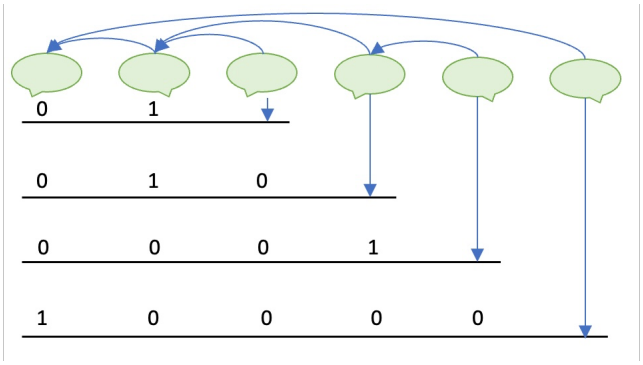

Figure 2: Illustration of structure pretraining. Utterance bubbles denote an utterance sequence with the "reply_to" relation between them. The labels represent the ground truth for each corresponding subsequence.

reduce computational cost, instead of applying a full transformer model, we use only four transformer layers.

3.4 Multitask Training with Distance to Derailment

Following previous work (Chang and Danescu-Niculescu-Mizil 2019), we simulate how conversations evolve in reality and predict conversation derailment dynamically. At inference time, the model is fed with the sequence $\{u_1, u_2, \ldots, u_{n-1}\}$ utterance-by-utterance and makes a prediction at each step. The model stops whenever a positive prediction is made, which indicates that the conversation is expected to derail. At training time, however, Chang and Danescu-Niculescu-Mizil (2019) apply a static training strategy, where the model is trained only with full sequences up to the derailing utterance $\{u_1, u_2, \ldots, u_{n-1}\}$. We posit that this discrepancy of feeding utterances between training and inference may bias the model to overestimate the probability of derailment for long inputs. Therefore, we propose to unify training and inference with the same dynamic strategy by adopting a regression task besides the binary classification task.

Suppose our sample pair is $(\{u_1, u_2, u_3, u_4\}, \text{positive})$, $(\{u'_1, u'_2, u'_3, u'_4\}, \text{negative})$. We observe that the distance from each civil utterance to the derailing utterance could provide additional cues for the model to learn. By predicting the distance to derail, another benefit is that we can expand the training set by a factor of the average conversation length.

For the positive sample, we can train on $(\{u_1\}, 3)$, $(\{u_1, u_2\}, 2)$, $(\{u_1, u_2, u_3\}, 1)$ for the regression task, where the targets $3, 2, 1$ represent the distance from the current sequence to the derailing utterance u_4. In the extreme, if the target is infinity, it means the conversation doesn't derail. In other words, a larger target suggests a lower chance of derailment.

Therefore, for the negative sample $(\{u'_1, u'_2, u'_3, u'_4\}, 0)$, we expand into $(\{u'_1\}, \infty)$, $(\{u'_1, u'_2\}, \infty)$, $(\{u'_1, u'_2, u'_3\}, \infty)$, $(\{u'_1, u'_2, u'_3, u'_4\}, \infty)$. A practical consideration is that if we set the target to be infinity

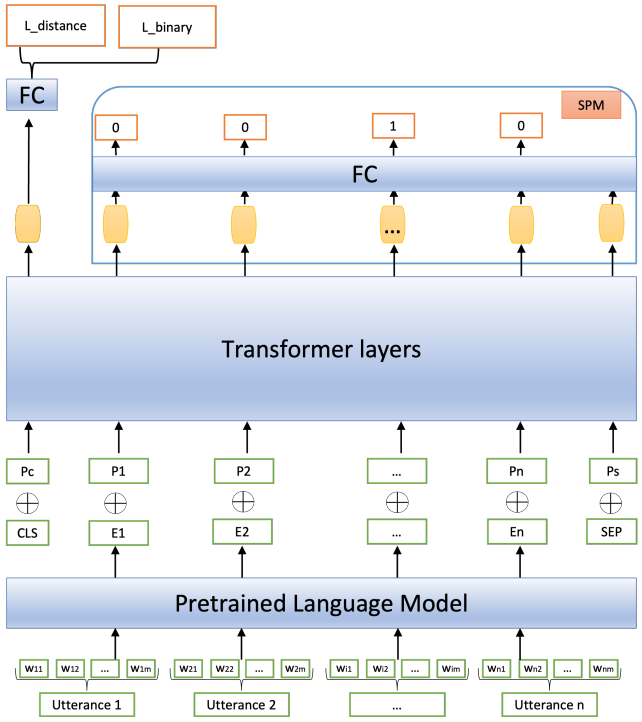

Figure 3: Illustration of a hierarchical transformer with a multitask learning scheme. The top-left component represents the multitask training scheme. SPM refers to the structural pretraining module.

for negative samples, the regression loss would be too large. Practically, we can simply set the longest length of conversation in the dataset as the target for each expanded negative sample. We add another regression head following the fully connected linear layer. Figure 3 illustrates the multitask training architecture on the top left. Our new loss function becomes

$$L = \alpha L_{distance} + (1 - \alpha) L_{binary},$$

where α controls the weight of each loss. We use mean squared error as the loss for $L_{distance}$ and cross entropy for L_{binary}.

3.5 Conversation Structure Pretraining

Conversations between a group of people on social media usually from a tree structure. Platforms such as Reddit, Wikipedia talk page, and Twitter have a clearly defined "reply-to" attribute for each comment. Previous research pays inadequate attention to the structure when modeling conversations. To investigate whether such structure is relevant in conversation development, we propose a scheme to pretrain our model on such structure in an unsupervised way. Specifically, we feed a sequence of utterance embeddings $\{E_1, E_2, \ldots, E_n\}$ to the second-level transformer model as usual, and denote the output from the last transformer layer as $\{O_1, O_2, \ldots, O_n\}$, which is followed by a fully connected layer with a softmax activation function. That is, we

obtain the following:

$$O_i = \text{Trans\_Layer}(E_i), i = 1, \ldots, n-1$$

$$p_i = \text{softmax}\left(\text{FC}(O_i)\right), i = 1, \ldots, n-1$$

and the loss function is

$$L = -\sum_{i=1}^{n-1} y_i \log p_i$$

where $y_i = 1$ if the ith utterance is the parent of the nth utterance.

Figure 2 illustrates the ground truth labels for all the subsequences of a conversation. We refer to the pretraining module as the structural pretraining module (SPM).

3.6 Data Augmentation Strategy

One advantage of multitask training and conversation structure pretraining is that the dataset can be augmented by a factor of the average length of the conversations, as explained in Section 3.4. This straightforward strategy, however, may not produce the best results. After examining some conversations, we observe that for the longer conversations, the first few utterances provide weak indications of derailment, which may confuse the model.

We assume that the derailment is the result of the cumulative effect of multiple exchanges. Therefore, we apply a different augmentation strategy to keep at least half of the previous utterances in the conversation. Suppose the length of the conversation is n, we start expanding the data at length $i = \lfloor n/2 \rfloor$. Now each data sample is extended to a sequence of samples $\{u_1, \ldots, u_i\}, \{u_1, \ldots, u_i, u_{i+1}\}, \ldots, \{u_1, \ldots, u_i, u_{i+1}, \ldots, u_{n-1}\}$. Our results show that this strategy works better.

4 Experiments and Results

We evaluate our model on two canonical conversation derailment datasets proposed in previous research (Zhang et al. 2018; Chang and Danescu-Niculescu-Mizil 2019). Both datasets reflect the same philosophy with respect to collecting the data. The essential idea in both is to construct (1) positive samples, where the first few utterances are civil but eventually develop into a personal attack and (2) negative samples, where the whole conversation is civil. Another consideration is to avoid potential bias related to the conversation topic. For example, if the topic distribution between negative and positive samples is different, the model is likely to capture that fact and thus predict a high probability of derailment for conversations whose topics are similar to those in the positive samples. One way to counter this challenge is to make sure that every positive sample has a corresponding negative sample where the conversations cover similar topics. Therefore, the procedure has two steps. First, identify conversations that contain personal attacks. Second, for each derailing conversation, collect a civil conversation that covers the same topic. As a result of this procedure, the dataset is balanced between derailing and civil conversations.

4.1 Datasets

Wikipedia talk page (WTP) The WTP dataset was introduced by (Zhang et al. 2018) and then expanded by (Chang and Danescu-Niculescu-Mizil 2019) using the same procedure. Every Wikipedia article is associated with a Talk Page, where Wikipedia editors discuss its editing. Each page usually has multiple sections focusing on the discussion of different editing problems. Every section is in the form of a conversation. The goal is to select conversations that start out as civil but derail into personal attacks afterwards. Wikipedia contains millions of pages and conversations. To alleviate the effort of manually going through all conversations, we applied a toxicity classifier to select candidate conversations that contain toxic utterances. Toxicity, however, is not always equivalent to personal attacks. Therefore, after selecting the candidates, we manually selected conversations with personal attacks.

The classifier used is provided by Perspective API, which is trained on Wikipedia talk page comments that have been annotated by humans. The classifier provides a toxicity score ranging from 0 to 1 for each utterance. Two types of conversation are preselected: (a) those that are civil throughout—all comments in the conversation have a toxicity score below 0.4 and (b) those that are civil for the first exchange (two comments), but turn toxic afterwards—there is a comment with toxicity score above 0.6. These two numbers are chosen empirically by running examples with the API. When the score is lower than 0.4, it has high fidelity that it is civil and when it is over 0.6, it is toxic. However, the exact numbers do not matter because we use the classifier only as a first step to identify candidate conversations.

As explained above, to avoid the model from capturing spurious correlations, such as conversation topics or length, each positive sample (conversations starting out civil and ending with personal attacks, i.e., the candidates from the first step) is paired with a negative sample (conversations are civil throughout). These negative samples are selected as follows: (1) they are from the same Wikipedia talk page so their discussions are about the same article; (2) they have similar lengths; and (3) they take place close in time. We refer the reader to Section 3 of the original paper (Zhang et al. 2018) for additional details. This procedure produces a dataset that contains 2,094 pairs of conversations, split into 60%–20%–20% segments. Formally, we denote 1 as a personal attack comment and 0 as a civil comment. Then, a sample data pair can be represented as $(0, \ldots, 0), (0, \ldots, 0, 1)$. As a result of the procedure, the dataset has the same number of positive and negative samples.

Reddit ChangeMyView (CMV) The Reddit CMV dataset was crafted by (Chang and Danescu-Niculescu-Mizil 2019). ChangeMyView is a subreddit where a reddit users may post their view on some topic and challenge other users to give them a reason to change their view. Rule 2 of the ChangeMyView subreddit states "Don't be rude or hostile to other users." The platform moderator may delete any comment that appears to include a personal attack and replace the comment with the word "deleted." Then, all previous comments up to the "deleted" comment make

up a positive conversation sample with derailment at the end. The deleted comment is not visible. However, as the setting is to predict whether a conversation will derail or not, rather than to classify an existing derailing utterance, the content of the derailing utterance is not needed for training. To apply the topic and length control pairing, as in the WTP datasets, each positive and negative pair of similar length is chosen from the same top-level post. An additional control is applied to select conversations where the deleted comment is from a user who previously participated in the conversation. We refer the reader to Section 3 of the original paper (Chang and Danescu-Niculescu-Mizil 2019) for additional details. This procedure produces a dataset of 3,421 pairs of conversations, which is also split into 60%–20%–20% parts. It is worth noting that there is no post hoc annotation and examination of this dataset. Therefore, there is no guarantee that for either positive or negative conversations, all comments prior to the last one would be civil, e.g., if the moderator missed deleting them. Thus, the CMV dataset may contain more noise than the WTP dataset.

4.2 Experimental Setup

Baseline We compare our approach with a few important works from previous research. Specifically, We compare the performance with a straightforward bag-of-words model, which simply concatenates all utterances and converts them into a bag-of-words vector. We also compare with the CRAFT model proposed by the original paper (Chang and Danescu-Niculescu-Mizil 2019). In addition, we consider two recent works from (Janiszewski, Lango, and Stefanowski 2021) and (Kementchedjhieva and Søgaard 2021), which are evaluated on the same dataset.

Training process To facilitate the learning process, we experimented with different configurations. We implemented our model with the HuggingFace library and set the learning rate to be 1×10^{-5} with a batch size of 32 for the multitask learning part. We experiment with different α values and set it to be 0.3 for best performance. The experiment shows that we need to be conservative with the regression task.

The model was trained with the Adam optimization algorithm. We observe that using only the first or second utterance as inputs would have a negative impact on performance, possibly due to the fact that the first and second utterances do not provide enough cues for the model to predict derailment. Therefore, we adopt the strategy described in Section 3.6. With this strategy, the target for regression task is in the range of 0 to 5. During training, we evaluate the performance on the dev split every 100 iterations and save the checkpoint only if the performance of the current iteration is better than that of the last checkpoint. For the SPM adaptive pretraining module, as the goal of the training is to enable the model to be able to capture general structural information existing in the conversation, we adopt the full data augmentation strategy as described in Section 3.4.

Evaluation Following Janiszewski, Lango, and Stefanowski (2021), we evaluated the performance of the model with respect to these metrics: accuracy, precision, recall, FPR (false positive rate), and F1 score. To have a fair comparison, during evaluation, we looked only at the outcome from the binary prediction head and ignored the prediction from the distance-to-break head. The evaluation was done in a progressive manner. That is, for a conversation of length n, whenever the model makes a positive prediction when feeding with utterance sequence $(\{u_1\}, \{u_1, u_2\}, \ldots, \{u_1, u_2, \ldots, u_{n-1}\})$, it counts as a derailment prediction. If the entire preceding sequence is predicted as negative, the conversation is deemed to be not derailed.

4.3 Results and Analysis

Table 1 shows the results of our various model variants, together with the models from previous research. It is worth noting that the performance of CRAFT in (Janiszewski, Lango, and Stefanowski 2021) and (Kementchedjhieva and Søgaard 2021) are a bit lower than what the original paper reported, due to variations during training. We cite the results from (Kementchedjhieva and Søgaard 2021) which has taken the average of ten runs of CRAFT. The BERT model in the table is the architecture from (Kementchedjhieva and Søgaard 2021), which simply add a classification head on top of the original BERT model. We follow the static training details described by the paper and report the results we have. Hierarchical-Base is our base model of the hierarchical transformer, which consists of only one binary prediction head without data augmentation. Hierarchical-Multi is the multitask learning model, which has two prediction heads and uses data augmentation during training. Hierarchical-Multi+Pretrain includes pretraining with structural information followed by the Hierarchical-Multi model. Overall, our models achieve better performance regarding most of the metrics. Between our own variants, the result are mixed.

We also experiment with different values of α with Hierarchical-Multi on the WTP dataset to investigate the optimal weight of each task contributing to the loss. Fig 4 shows that the optimal α should be 0.3. The insight is that we should be fairly conservative with respect to the regression task. This observation echoes the fact that we should not augment the data to the full length of each conversation. The model learns better when more utterances are being observed.

A major difference between our architecture and previous models is that we leverage the power of pretrained language models and self-attention mechanisms. Both CRAFT (Chang and Danescu-Niculescu-Mizil 2019) and HRED (Janiszewski, Lango, and Stefanowski 2021) consist of two LSTM layers, and they pretrained the model over 1 million conversations with an autoregressive language model objective to get the best performance, which is a heavy cost and requires gathering a large amount of data from the same domain. It's also not easy to adapt the pretrained language model to other tasks. As we can see, all model variants based on pretrained language models achieve better F1 scores on both the WTP and CMV datasets. HRED has the lowest FPR on the WTP dataset while CRAFT has the highest recall on the same dataset. For all other metrics on both datasets, our models have higher performance. The BERT model is from

Model	Wikipedia talk pages					Reddit CMV				
	Accuracy	Precision	Recall	FPR	F1	Accuracy	Precision	Recall	FPR	F1
BoW	56.5	55.6	65.5	52.4	60.1	52.1	51.8	61.3	57.0	56.1
CRAFT	64.4	62.7	71.7	-	66.9	60.5	57.5	**81.3**	-	67.3
HRED	63.9	63.8	64.1	**36.2**	64.0	55.6	54.6	65.8	54.7	59.7
BERT	63.8	61.1	75.7	48.1	67.6	65.7	64.4	70.2	38.8	67.1
Hierarchical-Base	62.9	60.3	75.2	49.5	66.9	64.3	**67.1**	56.2	**27.6**	61.1
Hierarchical-Multi	65.2	62.3	**76.9**	46.4	**68.9**	64.2	62.0	73.8	45.2	**67.4**
Hierarchical-Multi+Pretrain	**65.2**	**64.2**	69.1	38.6	66.5	**66.2**	66.5	65.2	32.8	65.9

Table 1: Results of the proposed model on two datasets, compared to three previous approaches (BoW and CRAFT (Chang and Danescu-Niculescu-Mizil 2019), HRED (Janiszewski, Lango, and Stefanowski 2021)), BERT (Kementchedjhieva and Søgaard 2021). Hierarchical-Base (base version of the hierarchical transformer), Hierarchical-Multi (hierarchical transformer with multiple learning scheme), Hierarchical-Multi+Pretrain (hierarchical transformer with multiple learning scheme and pretraining module). Section 2.3 introduces the last four methods shown above.

(Kementchedjhieva and Søgaard 2021), which has a classification head on top of BERT. Even though with a simple BERT model, the performance is still competitive. Therefore, we answer **RQ1** postively.

Comparing the three different variants of our hierarchical model, the multitask learning architecture yields the best F1 scores among all models on both datasets, which demonstrates the effectiveness of our multitask learning and data augmentation strategy. Observe from our experiment that, if we change our data augmentation strategy from ratio sampling to full sampling, the performance decreases, which indicates that the model needs more context to learn the factors leading to derailment. Conversely, the base version of hierarchical transformer has a lower FPR than Hierarchical-Multi on the CMV dataset, which indicates that the distance-to-derailment information encourages the model to make positive predictions. This is echoed by the fact that the Hierarchical-Multi model has a higher recall rate. Hierarchical-Multi has the highest F1 score on the WTP dataset. Both Hierarchical-Multi and Hierarchical-Multi+Pretrain have higher performance than Hierarchical-Base in terms of F1 score on the CMV datasets, which demonstrates the effectiveness of the multitask learning and data augmentation strategy. Therefore, we answer **RQ2** positively.

Contrary to our expectations, adding the reply-to pretraining process does not yield improved performance. There might be three reasons for this. First, we didn't pretrain our model on other data within the same domain, but limited our pretraining to within the two datasets. Although the dataset is augmented by a factor of half of the conversation length, the total number of data points is far less than the amount of data that is used by Chang and Danescu-Niculescu-Mizil (2019). Second, the reply-to relation between the utterances in these two datasets possibly doesn't align well with the semantics of the utterances. Third, derailment may not relate too much to the semantics of each utterance. We observe that most derailments happen due to impoliteness. The conversation topics in both datasets are diverse and the pretraining may cause the model to capture noise from the dataset. Consequently, we answer **RQ3** negatively.

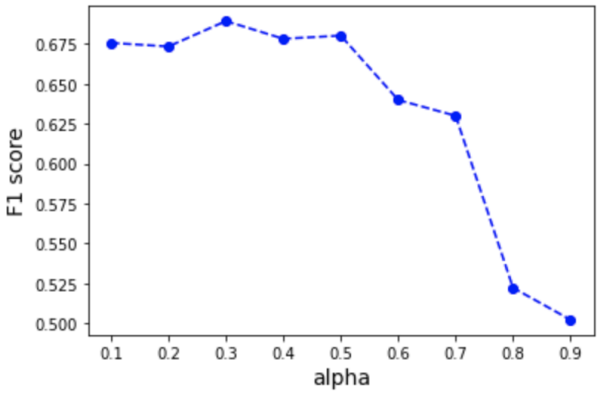

Figure 4: F1 score with respect to different α value on the WTP dataset

4.4 How Early is the Warning?

To investigate how early the warning is triggered for a conversation, we perform an analysis on the two datasets. Figure 5 shows the distribution of the percentage of utterances that have elapsed when the Hierarchical-Multi model makes a prediction of derailment for all the positive samples in the testing set. We observe that around 80% of warnings are issued when fewer than five utterances have been seen by the model and the average length of the derailing conversations in the testing set is 7.2. We also show the average number of utterances between the issuance of a warning and the derailing utterance for four models. The distance is around three for the WTP dataset and four for the CMV dataset. These figures generally match those in previous research (Kementchedjhieva and Søgaard 2021), (Chang and Danescu-Niculescu-Mizil 2019) have discovered.

5 Conclusion and Future Work

We focus on a new intervention perspective regarding detecting and moderating toxic and abusive behaviors in online forum conversations. Rather than predicting whether a conversation contains toxic content retrospectively, we seek to

	WTP	CMV
BERT	2.73	3.98
Hierarchical-Base	2.96	3.94
Hierarchical-Multi	2.86	3.45
Hierarchical-Multi+Pretrain	2.98	3.78

Table 2: Mean number of utterances between the issue of warning and the derailing utterance. BERT(Kementchedjhieva and Søgaard 2021). Hierarchical-Base (base version of the hierarchical transformer), Hierarchical-Multi (hierarchical transformer with multiple learning scheme), Hierarchical-Multi+Pretrain (hierarchical transformer with multiple learning scheme and pretraining module).

predict whether an ongoing conversation will break down. We propose a hierarchical transformer architecture to capture both utterance-level and conversation-level semantics leveraging the power of pretrained language models. In addition, we propose new ways to integrate conversational structure and the distance-to-derailment information and achieve better F1 scores than previous approaches on two canonical conversation derailment datasets.

Although our model mainly addresses the problem of predicting conversation derailment dynamically, it is a general approach for conversation modeling and can be adapted to address other conversation prediction tasks. For example, to predict whether a goal-oriented conversation would succeed in the end, we can set the label for successful conversations as 1 and unsuccessful conversation as 0. In addition to a simple binary prediction, our distance-to-derailment prediction could provide extra time-sensitive information. For tasks where the inherent conversation structure matters, our model provides a natural way to exploit such structural information. We defer the application of our approach on other conversation prediction tasks to future work.

With all the strengths that come with our approach, we have identified some important limitations and directions for future research. First, we have limited our reply-to structure pretraining to the two datasets we evaluate, which doesn't seem to provide enough data for this kind of pretraining. Therefore, to leverage the structure information, we could continue to train the model on any conversations that happened within the same forum. Second, we can adopt specialized architectures that align well with structural data, such as graph neural networks, which are specifically designed for capturing tree-like structure information. Third, a promising direction is to capture speaker identity or explore how moral postures (Xi and Singh 2023) affect derailment prediction. For most conversations, there might be a pattern where certain types of users are more inclined to attack other people. The ability to model different speaker types may be important in such kinds of prediction.

6 Ethical Impacts

Our work proposes ways to model conversations and predict whether a conversation will develop into a personal attack. The datasets we work on contain only user IDs from Reddit and Wikipedia talk page, which do not reveal personal identities (though if a user has revealed their identity in some other post, it could be discovered). The proposed framework and trained model can be applied by social platforms to assist with content moderation, where early warnings can be issued to prevent personal attacks from happening. As with many other pretrained deep learning models, our model could be exploited by users so that they learn the pattern to avoid the censor. Our model could also be limited in its ability to accurately capture conversation dynamics as the domain and topics evolve over time. However, we posit that active and continual learning could help mitigate this problem.

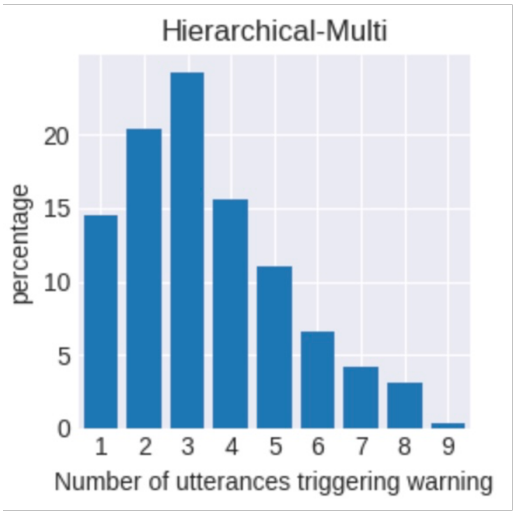

Figure 5: Percentage of the number of utterances elapsed when the model makes a positive prediction of derailment.

Acknowledgments

Thanks the anonymous reviewers for their comments and to the US National Science Foundation (grant IIS-2116751) for partial support.

References

Bao, J.; Wu, J.; Zhang, Y.; Chandrasekharan, E.; and Jurgens, D. 2021. Conversations Gone Alright: Quantifying and Predicting Prosocial Outcomes in Online Conversations. In *Proceedings of the Web Conference 2021*, WWW '21, 1134–1145. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383127.

Chandrasekharan, E.; Samory, M.; Srinivasan, A.; and Gilbert, E. 2017. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 3175–3187. New York: Association for Computing Machinery.

Chang, J. P.; and Danescu-Niculescu-Mizil, C. 2019. Trouble on the Horizon: Forecasting the Derailment of Online

Conversations as they Develop. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4743–4754. Hong Kong, China: Association for Computational Linguistics.

Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2015. Antisocial Behavior in Online Discussion Communities. In *International AAAI Conference on Web and Social Media (ICWSM)*, 61–70. Association for the Advancement of Artificial Intelligence.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Field, A.; Blodgett, S. L.; Waseem, Z.; and Tsvetkov, Y. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1905–1925. Online: Association for Computational Linguistics.

Fortuna, P.; and Nunes, S. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Survey*, 51(4).

Guo, Z.; and Singh, M. P. 2023. Representing and Determining Argumentative Relevance in Online Discussions: A General Approach. In *Proceedings of the 17th International AAAI Conference on Web and Social Media (ICWSM)*, 1–8. Limassol, Cyprus: AAAI Press. To appear.

Guo, Z.; Zhang, Z.; and Singh, M. P. 2020. In Opinion Holders' Shoes: Modeling Cumulative Influence for View Change in Online Argumentation. In Huang, Y.; King, I.; Liu, T.; and van Steen, M., eds., *WWW: The Web Conference, Taipei, Taiwan, April 20-24, 2020*, 2388–2399. Proceedings of The International World Wide Web Conference 2020.

Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8342–8360. Online: Association for Computational Linguistics.

Hessel, J.; and Lee, L. 2019. Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1648–1659. Minneapolis, Minnesota: Association for Computational Linguistics.

Howard, J.; and Ruder, S. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 328–339. Melbourne, Australia: Association for Computational Linguistics.

Hua, Y.; Danescu-Niculescu-Mizil, C.; Taraborelli, D.; Thain, N.; Sorensen, J.; and Dixon, L. 2018. WikiConv: A Corpus of the Complete Conversational History of a Large Online Collaborative Community. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2818–2823. Brussels, Belgium: Association for Computational Linguistics.

Ive, J.; Anuchitanukul, A.; and Specia, L. 2021. Revisiting Contextual Toxicity Detection in Conversations. *ArXiv*, abs/2111.12447.

Janiszewski, P.; Lango, M.; and Stefanowski, J. 2021. Time Aspect in Making an Actionable Prediction of a Conversation Breakdown. In Dong, Y.; Kourtellis, N.; Hammer, B.; and Lozano, J. A., eds., *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*, 351–364. Cham: Springer International Publishing.

Kementchedjhieva, Y.; and Søgaard, A. 2021. Dynamic Forecasting of Conversation Derailment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7915–7919. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Khanpour, H.; Guntakandla, N.; and Nielsen, R. 2016. Dialogue Act Classification in Domain-Independent Conversations Using a Deep Recurrent Neural Network. In *Proceedings of COLING, the 26th International Conference on Computational Linguistics: Technical Papers*, 2012–2021. Osaka, Japan: The COLING 2016 Organizing Committee.

Kumar, S.; Cheng, J.; and Leskovec, J. 2017. Antisocial Behavior on the Web: Characterization and Detection. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW Companion, 947–950. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints*, arXiv:1907.11692.

Meaney, J.; Wilson, S.; Chiruzzo, L.; Lopez, A.; and Magdy, W. 2021. SemEval 2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval 2021)*, 105–119. Association for Computational Linguistics (ACL). 15th International Workshop on Semantic Evaluation, SemEval 2021 ; Conference date: 05-08-2021 Through 06-08-2021.

Mojica, L. G. 2017. A Trolling Hierarchy in Social Media and A Conditional Random Field For Trolling Detection. *The Computing Research Repository*, abs/1704.02385.

Mozafari, M.; Farahbakhsh, R.; and Crespi, N. 2019. A BERT-based transfer learning approach for hate speech detection in online social media. In *Complex Networks 2019: 8th International Conference on Complex Networks and their Applications*, volume Studies in Computational Intelligence book series (SCI, volume 881) of *Complex Networks*

and Their Applications VIII : Volume 1, Proceedings of the Eighth International Conference on Complex Networks and Their Applications, 928–940. Lisbonne, Portugal: Springer.

Pavlopoulos, J.; Sorensen, J.; Dixon, L.; Thain, N.; and Androutsopoulos, I. 2020. Toxicity Detection: Does Context Really Matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics.

Sordoni, A.; Bengio, Y.; Vahabi, H.; Lioma, C.; Grue Simonsen, J.; and Nie, J.-Y. 2015. A Hierarchical Recurrent Encoder-Decoder for Generative Context-Aware Query Suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM, 553–562. New York: Association for Computing Machinery. ISBN 9781450337946.

Vidgen, B.; Harris, A.; Nguyen, D.; Tromble, R.; Hale, S.; and Margetts, H. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, 80–93. Florence, Italy: Association for Computational Linguistics.

Wang, X.; Shi, W.; Kim, R.; Oh, Y.; Yang, S.; Zhang, J.; and Yu, Z. 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*, 5635–5649. Florence, Italy: Association for Computational Linguistics.

Xi, R.; and Singh, M. P. 2023. The Blame Game: Understanding Blame Assignment in Social Media. *IEEE Transactions on Computational Social Systems*, 10: 1–10.

Zhang, J.; Chang, J.; Danescu-Niculescu-Mizil, C.; Dixon, L.; Hua, Y.; Taraborelli, D.; and Thain, N. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1350–1361. Melbourne, Australia: Association for Computational Linguistics.