

A Theoretically Grounded Approach to Summarizing Conversation Dynamics for Forecasting the Derailment of Online Conversations

Yingxue Fu, Anaïs Ollagnier

Universite Cote d'Azur, CNRS, Inria, I3S

Sophia Antipolis, France

fuyingxue321@gmail.com

Abstract

Conversation derailment prediction represents a new paradigm of toxicity detection, where a system predicts from the start of a conversation whether it will derail into toxic exchanges, allowing moderators and users to act preemptively before harm is done. This approach requires a deep understanding of conversation dynamics. Previous work relies on linguistic features rooted in linguistic and social theories. While these features provide signals of conversation dynamics, they are exploratory in nature and potentially reflect a fraction of the overall pragmatic devices shaping the conversation trajectory. To capture the pragmatic dimension of conversations systematically, we start with a framework for annotating pragmatic information of conversations systematically and design summary generation methods to capture conversation trajectory dynamically. We achieve about 10% performance increase over a simple baseline, and 6.47% increase over a strong baseline on a dataset, and a slight performance increase on a benchmark dataset for the task of summarizing conversation dynamics.

1 Introduction

Online discussions involve participants from diverse backgrounds, and it is not rare that they derail into toxic exchanges. Anti-social behaviors (ASBs) online, such as hate speech (ElSherief et al., 2021; Ocampo et al., 2023), trolling (Lee et al., 2022) and more generally, toxicity (Pavlopoulos et al., 2020), pose a threat to maintaining inclusive and efficient online discussion platforms. The dominant approach to combating ASBs is *reactive* moderation, in which volunteers from within the community or professional moderators employed by platform operators monitor online discussions and ban some participants or remove comments that violate community policies (Chang et al., 2022). Owing to the large data volume, automatic systems can be

helpful to alleviate moderators' burdens. Substantial research efforts have been devoted to developing NLP systems for this purpose (Davidson et al., 2017; Waseem et al., 2017; Wulczyn et al., 2017; Pavlopoulos et al., 2017; Gao and Huang, 2017; Zampieri et al., 2019; Chakravarthi et al., 2024; Jin et al., 2025).

However, such an endeavor has been shown to be both technically and ethically challenging (Jurgens et al., 2019; Waseem and Hovy, 2016; Fortuna et al., 2020), with a growing number of benchmarks and datasets created to merely capture the linguistic and social nuances of ASBs (ElSherief et al., 2021; Sap et al., 2020; Poletto et al., 2021). As platforms struggle to scale moderation and accountability, there is a growing need for *predictive* systems that can forecast harmful interactions before they happen (Ollagnier, 2026; Zhang et al., 2018; Chang and Danescu-Niculescu-Mizil, 2019; Kementchedjheva and Søggaard, 2021; Altarawneh et al., 2023; Zhang et al., 2025). Such systems enable more effective interventions before toxicity spreads or escalates, thus mitigating potentially severe consequences for individuals and communities. Moreover, the *post-hoc* reactive paradigm ignores the fact that uncivil behavior in discussion platforms is not solely the product of “bad actors” but can often emerge from ordinary users when they happen to be involved in heated discussions (Chang et al., 2022). Predictive systems can be used to inform conversation participants of the risks of conversation derailment so that they can avoid escalating tense discussions proactively.

Consider the example in Table 1. A reactive system will classify whether a message is toxic after it appears. In the example, it is likely that messages (1)-(6) are labeled as non-toxic and message (7) is flagged as toxic. In contrast, a predictive system anticipates whether the conversation will eventually become toxic as the conversation develops. Therefore, it might forecast a risk that the conversation

(1) A. I was on break for 5 months and also out of Germany. The Tk is demanding me money for whole 5 months. As my financial condition is not so good being student, can Tk refund me the money or can I get a tax return?
(2) B. Were you a registered student at the time?
(3) A. Yeah.
(4) B. Then you were obliged to have insurance. If you do not do this, customs will sooner or later collect the money by means of seizure.
(5) A. How would they take money by seizure when I am broke? Or do I have to sell my body parts to pay them?
(6) B. They will block your bank account.
(7) A. They didn't blocked my bank account, the only thing they did is block my enrollment letter.
You don't know jackshit about German administration and bureaucracy. Stop commenting half ass confirmed information out here.

Table 1: A conversation taken directly from the subreddit *AskGermany*, originally extracted from the Pushshift repository (Baumgartner et al., 2020), hence containing grammatical errors. It eventually derails into a personal attack. Some comments are omitted here, without affecting the overall tone.

will turn toxic at message (5) or even earlier. Moderators can intervene at that point, or alternatively, a system can prompt Participant A to rephrase the message. As is clear, an important difference between the two approaches is that in the predictive paradigm, all the messages of a conversation have the same label, i.e., whether the conversation finally goes toxic or not, while in the reactive paradigm, each message has its own label of toxicity.

A predictive system needs to capture not only the content of a message, but also the conversation dynamics or trajectory in order to anticipate the future development of the conversation. Previous studies (Zhang et al., 2018) employ politeness strategies, signaled by lexical markers, such as *please*, and rhetorical roles of comments, such as factual check or expressing an opinion, to capture pragmatic devices in exchanges. Social orientation labels are used in the study by Zhang et al. (2025) to model conversation dynamics, which characterize interpersonal interactions through a set of core dimensions, such as power, benevolence (warmth and positivity of the interaction), arousal (the level of energy or excitement expressed), and political inclination. Hua et al. (2024) propose a summarization method for capturing conversation dynamics. An effective summarization system must select the most relevant information while keeping the summary concise. While these social and linguistic cues provide useful insights for forecasting toxicity of conversations, there lacks a systematic approach

to capturing conversation dynamics. This is one of the reasons why Hua et al. (2024) introduce a multi-step procedure for human annotators and draw inspiration from the process of human annotation to design prompt templates for use on large language models (LLMs).

In this study, we build on the work by Hua et al. (2024) and adopt a theoretically grounded framework for systematically capturing pragmatics of online exchanges and summarizing conversation trajectories. We adopt the framework of annotating pragmatic information developed by Fu and Ollagnier (2025) and configure a summarization generation module to capture conversation dynamics.

Our contributions include: 1) We propose a new method for capturing conversation dynamics. Experimental results show that automatic annotation of pragmatic information is sufficient to achieve high performance, without the need for manual annotation. 2) Our experimental results show that our approach achieves about a 10% increase in conversation derailment prediction over the baseline where the original conversation messages are used for summarizing conversation dynamics. In addition, we compare our method with that proposed by Hua et al. (2024), and our approach achieves a 6.74% increase on a dataset and slightly higher performance on predicting conversation derailment dynamically on the benchmark dataset created by Hua et al. (2024). 3) We perform ablation studies and show insights into how the proposed ways of summarizing conversation dynamics differ from each other.

2 Related Work

Toxicity is defined as “a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion” (Wulczyn et al., 2017). It is understood as a general term covering a spectrum of antisocial behaviors online, such as cyberbullying (Ollagnier et al., 2022; Verma et al., 2022; Menini et al., 2019; Sprugnoli et al., 2018), trolling (Mojica de la Vega and Ng, 2018; Lee et al., 2022, 2024; Miao et al., 2020), profanity (Nozza and Hovy, 2023), hate speech (Schmidt and Wiegand, 2017; Yang et al., 2023; Yoder et al., 2022), offensive language (Waseem et al., 2017; Deng et al., 2022) and so on. Dominant NLP research on this topic adopts a post-hoc reactive approach (Gao and Huang, 2017; Qian et al., 2019). However, a

new predictive paradigm attracts increasing attention, which was introduced by Zhang et al. (2018). With this approach, systems predict from the very start of a conversation whether the interaction between participants will become toxic/aggressive, thus allowing moderators to intervene preemptively before toxic exchanges appear.

Capturing the pragmatic devices of a conversation is essential for this task (Zhang et al., 2018, 2025; Chen and Wang, 2025). Linguistic markers for politeness strategies and types of rhetorical prompts used to initiate different types of conversations are shown to be indicative of future conversation derailment (Zhang et al., 2018). Social orientation labels, measured by a set of dimensions such as power, emotional intensity/positivity, and political inclination, are employed in the research by Zhang et al. (2025). User information and public perception of a comment, measured by subtracting its up-votes by down-votes, are used in the experiment by Altarawneh et al. (2023). These features are extracted using rule-based methods (politeness strategies), unsupervised clustering (types of rhetorical prompts), annotation with an LLM (social orientation labels) and metadata from the corpus (user information and public perception).

Another approach relies on conversation modeling to capture conversation dynamics (Kementchedjhieva and Sjøgaard, 2021; Chang and Danescu-Niculescu-Mizil, 2019; Altarawneh et al., 2023; Yuan and Singh, 2023). In this case, implicit knowledge about conversation turns and their relations is represented in the generated conversation flow. A separate module is generally trained with supervised learning to make use of such knowledge for forecasting conversation derailment.

Two ways for model training can be identified (Kementchedjhieva and Sjøgaard, 2021): static training and dynamic training. Static training features feeding the $N-1$ turns before the last turn N to the model and predicting whether the message at turn N is toxic or not (a conversation is generally truncated until the last turn where toxicity occurs). During inference time, the model is tested dynamically, where prediction is made at each turn about whether the conversation will become aggressive in the end. The predicted label is obtained by taking the one with the highest probability from the predictions for all the turns before the last turn of a conversation in the test set. This paradigm is adopted by Altarawneh et al. (2023), Yuan and Singh (2023) and Hua et al. (2024).

In contrast, dynamic training closes the gap between training and inference time. During training, a conversation is unfolded into multiple samples, each representing a different stage of the conversation, with all the samples having the same label as the last turn. This is the approach employed by Kementchedjhieva and Sjøgaard (2021) and Chang and Danescu-Niculescu-Mizil (2019). Kementchedjhieva and Sjøgaard (2021) propose to unroll only the last K turns as they believe that the first $K-1$ turns do not provide strong signals.

Most of the existing studies focus on the Conversations Gone Awry (CGA) dataset (Zhang et al., 2018), which is built on Wikipedia’s talk page discussions from the WikiConv dataset (Hua et al., 2018) and annotated manually, and Reddit Change-MyView (CMV) dataset (Chang and Danescu-Niculescu-Mizil, 2019), which is labeled based on clues from moderators’ delete operations, forming a potentially noisy dataset.

Despite existing efforts to tackle the challenging task, the linguistic features rooted in linguistic and social theories are exploratory in nature and possibly, they belong to only a fraction of pragmatic devices. Hua et al. (2024) experiment with summarization to address the problem. Similar to the conversation modeling approach, knowledge about conversation dynamics is implicit, and we still lack a systematic approach to unveiling the pragmatics of conversations to better capture conversation dynamics.

3 Proposed Method

Similar to Hua et al. (2024), we use summarization as an intermediate step to capture conversation dynamics, exploiting the impressive power of LLMs on various NLP tasks (Achiam et al., 2023; OpenAI, 2025). However, we focus on predicting conversation derailment *dynamically*, in line with the paradigm studied by Kementchedjhieva and Sjøgaard (2021) and Chang and Danescu-Niculescu-Mizil (2019).

3.1 Framework for Capturing Pragmatic Information

We adopt the framework proposed by Fu and Ollagnier (2025) for systematically capturing pragmatic information of conversations. This framework is not specific to any dataset and has high portability.

The annotation scheme is shown in Table 2.

Tag	Explanation	Categories
inferences	content of pragmatic inferences	free-text
	illocutionary act types	representatives, expressives, directives, commissives, declarations
most salient inference	the most salient inference	free text
as intended	whether the reply message agrees with the most salient inference of the parent message	yes/no
PRE/IMP	whether the most salient pragmatic inference belongs to presuppositions/implicatures	presupposition (PRE), implicature (IMP)
aggressive	toxicity of the message	yes/no

Table 2: The annotation scheme introduced by Fu and Ollagnier (2025).

It is grounded in Speech Act Theory developed by Austin (1962) and Searle (1975). Based on this theory, people use language to perform some acts, called speech acts. A speech act can be analyzed through three dimensions: the *locutionary act*, which means producing an utterance according to linguistic conventions; the *illocutionary act*, referring to performing an action in saying something, such as making a request, issuing a warning, or expressing a doubt; and the *perlocutionary act*, focusing on the effect the utterance has on the listener’s thoughts, feelings, or actions. For example, in *Could you close the window?*, the locutionary act is producing a grammatical English sentence, the illocutionary act is making a request, and the perlocutionary act is the listener responding by closing the window—or choosing not to (Fu and Ollagnier, 2025).

Each pragmatic inference can be categorized into one of five illocutionary act types following the taxonomy proposed by Searle (1975): *representatives*, which focus on the truth of a proposition; *expressives*, which convey psychological or emotional states; *directives*, which aim to prompt the hearer to act; *commissives*, which bind the speaker to a future action; and *declarations*, which bring about a new state of affairs through the act of making an utterance. These five types form an exclusive and exhaustive classification (Birner, 2025).

Among the potentially multiple pragmatic inferences, one may be more relevant to conversation participants. This phenomenon has been studied under Relevance Theory (Wilson and Sperber, 2004). In a conversation, the following message tends to focus on the most salient pragmatic inference, replying in agreement or disagreement, and the most salient message is not necessarily the same for the speaker and the respondent. The analysis of a speech act through the dimension of perlocutionary act is captured in the tag “most salient inference” and “as intended”.

The distinction between presuppositions/implicatures has been studied by Srikanth et al. (2024), who investigate the influence of these two types of pragmatic inferences on the development of healthcare dialogue systems. Implicatures refer to non-literal meaning of utterances (Grice, 1975) and presuppositions denote implicit assumptions that must hold for an utterance to be true. Srikanth et al. (2024) show that presuppositions require interlocutors to have the same assumptions, forming a stronger commitment to the implicit knowledge/beliefs underlying an utterance. In contrast, implicatures represent a softer way to express one’s assumptions.

With the annotation scheme, pragmatic inferences for each message are annotated first, among which the most salient inference is identified. Then, it is determined whether the most salient inference is shared by the interlocutors. As the final task is to predict toxicity of a conversation, disagreement between interlocutors might trigger tense discussions, possibly leading to toxic exchanges. The following step is to determine whether the most salient inference belongs to a presupposition (PRE) or an implicature (IMP).

3.2 Summary Generation

After pragmatic information is annotated, extracted and integrated with conversation messages, it is fed to a summarization module, which is configured to capture conversation dynamics. Inspired by the insights presented by Hua et al. (2024), we propose two approaches to generating summaries:

1. select and synthesize (select+synthesize): simply focusing on **selecting** and **synthesizing** the most relevant information for understanding the trajectory of the conversation.
2. incorporate **conversation dynamics** (CD): in addition to the instructions in se-

lect+synthesize, we ask the model to focus on *tones* in messages, such as ‘polite’, ‘rude’, ‘aggressive’, *tonal changes*, such as ‘tension maintained’, ‘escalated’, *patterns of interaction* that are indicative of future trajectories, such as ‘had a brief exchange’, *conversation strategies*, such as posing a question, and so on.

4 Implementation

We follow a two-step procedure: annotating pragmatic information and generating summaries for conversation derailment prediction. This setup allows us to better control the results of intermediate steps, and we leave exploration of end-to-end systems to future work.

For the sake of illustrating the setup, we highlight that the model makes predictions at each turn about whether the conversation will eventually go toxic using only information available up to that point. Crucially, when predicting at turn t (n turns in total for a conversation), the model may have access to:

- All messages at turns $1 \dots t$
- Pragmatic inferences for messages at turns $1 \dots t$
- Summaries derived only from messages at turns $1 \dots t$

No information from messages at turns $t+1 \dots n$ is accessible.

4.1 Annotation of Pragmatic Information

Given a dataset, we first annotate pragmatic inferences for each message in a conversation, following the framework by Fu and Ollagnier (2025). We specify that pragmatic inferences can be about sentiments, such as emotional state or stance towards some issues, and hidden knowledge from the message, including stereotypes and social constructs, pre-assumptions held by the author of the message, and implications derivable from the message. We stress that the model should avoid generating trivial inferences. For example, given a sentence *What kind of music should I play to my baby in the womb?*, inferences like the following are preferred over the ones shown with hash signs (Srikanth et al., 2024):

1. Babies can differentiate different sounds.

2. Certain kinds of music are more beneficial to babies in the womb than the others.

#Different kinds of music are available.

#Music is something that can be played.

As can be perceived, pragmatic information is annotated at the message level. However, for most_salient_inference and as_intended, knowledge about the immediately following message is required. Since LLMs can be used to perform the annotation task, when the annotator LLM is an auto-regressive model, it cannot foresee the incoming message, and the values for the two categories are likely to be low or determined based on surface heuristics. As these automatically generated annotations are provided as intermediate information to a summarization module, annotations of the two categories are kept in our experiments.

We use OpenAI GPT-4 (Achiam et al., 2023) for this step, making API calls to *gpt-4.1-2025-04-14*, which is the latest version of GPT-4 when we start our experiments. As shown by Fu and Ollagnier (2025), the annotations by GPT-4 are closer to manual annotations of pragmatic information in style¹. The maximum completion token is set to 4096, and the temperature is set to 0 for deterministic generation. The prompt template is shown in Appendix A. An annotated example is shown in Appendix B.

4.2 Extraction of Pragmatic Information

After the annotation step, messages and their corresponding annotated pragmatic information are extracted cumulatively, which means that for a conversation C consisting of n messages, predictions are made for (m_1) , $(m_1, m_2) \dots (m_1, \dots, m_n)$ one by one, where m denotes a message. Pragmatic information can be integrated with message information using predefined templates (Appendix C). Turn IDs and message authors, if given, are also extracted as a part of message information, but author IDs are to be anonymized with general references, such as “Speaker 1”.

To investigate the influence of different types of pragmatic information, we experiment with seven templates: 1) no pragmatic information is extracted, and summaries will be generated based on messages only (*msg_only*) (Appendix C.1); 2) only pragmatic information is used for summary generation, ignoring the original messages

¹Preliminary experiments suggest that the performance of Qwen, i.e., *Qwen2.5-1.5B-Instruct* (Qwen et al., 2025), is not acceptable for this task. One of the reasons is that it constantly skips messages.

(*prag_only*) (Appendix C.2); 3) all the annotated pragmatic information will be extracted, integrated with messages and used for summary generation (*msg_prag_full*) (Appendix C.3); 4) only the most salient pragmatic inference is used for summary generation, ignoring the original messages and other pragmatic inferences (*most_salient_only*) (Appendix C.4); 5) the original messages, pragmatic inferences and the most salient pragmatic inference are used for summary generation, omitting PRE/IMP and *as_intended* information (*msg_prag_salient*) (Appendix C.5); 6) the messages and the most salient inference are used for summary generation (*msg_salient*) (Appendix C.6); and 7) messages and pragmatic inferences are used for summary generation, omitting the most salient inference (*msg_prag*) (Appendix C.7).

4.3 Summary Generation

The second step is to generate summaries on the messages and extracted pragmatic inferences. The prompt templates for the two types of summaries discussed in section 3.2 are shown in Appendix D and Appendix E, respectively. Another LLM is used for this task. In our experiments, we prompt Qwen to generate summaries of conversation trajectories for each turn dynamically. Qwen represents an open-source LLM. We set *do_sample* to *False* and maximum number of tokens to 10,000, considering that conversations could be long in some cases.

We compare our method of summary generation with the three approaches studied by Hua et al. (2024): 1) procedural prompt, which is their best performing method, achieved with a prompt template designed based on human annotators' summarization procedure; 2) zero-shot prompt, which explicitly mentions generating summaries of conversation dynamics; and 3) traditional prompt, which simply focuses on the summarization task. The prompt templates used are shown in Appendix F, which are designed based on their released prompts². In this set of experiments, we only extract messages without using any pragmatic information, namely, the setting of *msg_only* discussed in section 4.2.

²https://github.com/CornellNLP/scd/blob/main/src/summary_generation_derailment_prediction/summ_gen.py

4.4 Datasets

The dataset released by Fu and Ollagnier (2025) is built on full reddit conversations. It contains annotations of pragmatic information by a human annotator and by GPT-4 and GPT-5 on the test set (31 conversations), and annotations by GPT-4 and GPT-5 on the training set (93 conversations). With manual annotations, we can study the influence of different types of pragmatic information on the conversation derailment prediction task with the seven templates discussed in section 4.2. As full threads contain interleaved comment and reply pairs, we first separate the threads into subthreads based on comment-reply relations, which are available from the dataset. Following the preprocessing step of the CGA dataset, the conversations formed by subthreads are checked to ensure that the first exchange is civil. As our task is conversation derailment *prediction*, conversations with fewer than three messages are discarded. After these filtering operations, 49 conversations are left in the test set. If a conversation contains a message labeled as toxic, the label for the whole conversation will be *toxic* and the conversation is kept up to the first occurrence of toxic messages. In evaluation, the output for the last message is ignored, since the task is *forecasting* the derailment of a conversation.

In addition, we perform experiments on the dataset created by Hua et al. (2024) to compare the performance of our methods for summary generation and their methods. The dataset is originally from the CMV dataset and it is balanced, containing pairs of derailing and non-derailing conversations under the same topic. The dataset contains 100 conversations in the test set (50 pairs), 234 conversations in the training set, and 100 conversations in the development set. Hua et al. (2024) work on truncated conversations, and summaries are generated for the whole conversation, instead of generated dynamically as the conversation develops. Four summaries are generated for each truncated conversation, and the final results are computed by averaging the classifiers' performance on them. For the task of capturing *future* trajectory, they remove the last three utterances from every conversation. As their interest is summarization, they only keep pairs where both conversations contain more than 10 utterances. The dataset does not contain annotations of pragmatic information. Therefore, we use GPT-4 to perform this step first (same settings as discussed in section 4.1).

5 Experimental Results

The experimental results on conversation derailment prediction with the proposed methods are shown in section 5.1, the results on extracting different types of pragmatic information are presented in section 5.2, and the results on the dataset by Hua et al. (2024) are discussed in section 5.3. Precision (P), recall (R) and F1 scores are shown in % and macro-averaging is used.

5.1 Conversation Derailment Prediction with Proposed Methods

Table 3 shows the results of conversation derailment prediction with the two summary generation methods (section 3.2). We prompt Qwen to generate summaries³ and predict conversation derailment on the dataset by Fu and Ollagnier (2025). Since the test set contains both manual and automatic annotations of pragmatic information, we show results for both cases. However, manual annotations of toxicity are used as gold labels throughout. Recall that in the case of `msg_only`, pragmatic information is not used. The prompt templates have been developed based on trials on the development set.

As is clear from the table, when the annotated pragmatic information is used, large improvements (10.96 for CD and 10.02 for `select+synthesize`) can be achieved over the baseline where only the conversation messages are used for generating summaries dynamically for toxicity prediction. The highest F1 scores (54.01 for CD and 51.91 for `select+synthesize`) are obtained when pragmatic information is annotated with GPT models, which suggests that manual annotation of pragmatic information is not necessary and automatic annotation is even preferred by the other LLM used for summary generation and toxicity classification⁴. When manual annotation of pragmatic information is used, the two summary generation methods do

³Only one summary is generated for each turn, since we focus on the dynamic paradigm of conversation derailment prediction, which is more costly than the scenario studied by Hua et al. (2024). Moreover, our experimental results on the dataset by Hua et al. (2024) suggest that standard deviation is small when multiple summaries are generated (section 5.3.2).

⁴We do not claim that automatically-generated annotations are superior to human annotations in an absolute sense. Rather, our claim concerns performance improvement: the gains we observe when using automated annotations may stem from greater consistency and scalability, or from the possibility that the downstream summarization model is better at processing automated annotations than the potentially noisier patterns of human annotations.

Pragmatic Information Extraction	SumGen Method	P	R	F1
<code>msg_only</code>	CD	49.86	49.89	43.05
<code>msg_only</code>	<code>select+synthesize</code>	49.66	49.74	41.89
<code>msg_prag_full</code>	CD (manual annotations)	54.32	54.32	50.81
<code>msg_prag_full</code>	<code>select+synthesize</code> (manual annotations)	54.95	54.88	50.86
<code>msg_prag_full</code>	CD (manualLabelGPT4Prag)	56.66	56.77	53.50
<code>msg_prag_full</code>	CD (manualLabelGPT5Prag)	59.02	58.66	54.01
<code>msg_prag_full</code>	<code>select+synthesize</code> (manualLabelGPT4Prag)	61.84	59.84	51.91
<code>msg_prag_full</code>	<code>select+synthesize</code> (manualLabelGPT5Prag)	54.54	54.51	49.71

Table 3: Results for the proposed methods compared with the baseline where messages are used directly for summary generation. *manualLabelGPT4Prag* refers to the setting that manual annotations of toxicity are used as target labels and GPT4 annotations of pragmatic information are used as features.

not differ much in performance, but when automatically annotated pragmatic information is used, higher performance is achieved with CD than `select+synthesize` ($\approx 2\%$ higher).

5.2 Results on Extracting Different Types of Pragmatic Information

Table 4 shows the results under different ways of extracting pragmatic information (section 4.2). Manual annotations of pragmatic information are used in this set of experiments and only a part of pragmatic information is extracted to test the influence when some pragmatic features are omitted. The baseline is using messages only for generating summaries, and `msg_prag_full` represents the case when all the annotated pragmatic features are extracted and integrated with messages for summary generation.

It can be seen that with CD, the lowest performance is achieved with `prag_only`. Recall that in this case, the original messages and annotations associated with the most salient inference are not used in the step of summary generation. In comparison, using messages together with the most salient inference alone achieves an F1 score (50.20) that is only slightly lower than using all the pragmatic features (50.81). This observation indicates that CD relies heavily on the most salient inference. With `select+synthesize`, the lowest F1 score is achieved

Pragmatic Information Extraction	SumGen Method	P	R	F1
prag_only	CD	46.73	46.81	43.22
most_salient_only	CD	52.84	52.87	49.72
msg_salient	CD	55.00	54.77	50.20
msg_prag	CD	52.82	52.79	48.91
msg_prag_salient	CD	54.98	54.49	48.94
msg_only	CD	49.86	49.89	43.05
msg_prag_full	CD	54.32	54.32	50.81
prag_only	select+ synthesize	59.55	59.16	54.39
most_salient_only	select+ synthesize	54.08	53.35	46.34
msg_salient	select+ synthesize	59.54	58.21	51.01
msg_prag	select+ synthesize	58.61	56.98	49.57
msg_prag_salient	select+ synthesize	58.17	57.99	53.28
msg_only	select+ synthesize	49.66	49.74	41.89
msg_prag_full	select+ synthesize	54.95	54.88	50.86

Table 4: Results for using different types of pragmatic information for summary generation with the two proposed methods. The results with msg_only and msg_prag_full are taken from Table 3 for reference.

with most_salient_only, where the original messages and other pragmatic inferences are omitted. This result may be related to the fact that the LLM for summarization and toxicity determination does not have sufficient information to select and synthesize the most relevant part to capture conversation trajectory. The F1 score achieved with prag_only is even higher than msg_prag_full, indicating the importance of sufficient underlying messages for this approach.

An example is shown in Appendix G to illustrate qualitatively the effect of incorporating pragmatic information in the task.

5.3 Comparison with Hua et al. (2024)

We distinguish between final results obtained by averaging results over all the turns (referred to as “all turns”), following previous work on the dynamic paradigm of conversation derailment prediction (Kementchedjheva and Søgaaard, 2021; Chang and Danescu-Niculescu-Mizil, 2019) and final results computed using only the prediction of the last turn of the conversation (denoted as “last turn only” for brevity), which is essentially the evaluation protocol by Hua et al. (2024).

Summary Type	P	R	F1
procedural prompt	50.74	50.75	47.54
zero-shot prompt	48.61	49.84	29.39
traditional prompt	59.85	55.57	43.86
CD (manualLabel-GPT5Prag)	59.02	58.66	54.01
select+synthesize (manualLabelGPT4Prag)	61.84	59.84	51.91
CD (msg_only)	49.86	49.89	43.05
select+synthesize (msg_only)	49.66	49.74	41.89

Table 5: Results for using the three approaches for summary generation studied by Hua et al. (2024). The last four rows present the performance obtained with our proposed methods in the case of msg_prag_full and in the case of msg_only, taken from Table 3 for comparison.

5.3.1 Results on the Dataset by Fu and Ollagnier (2025)

We apply the three approaches of summary generation studied by Hua et al. (2024), namely, procedural prompt, zero-shot prompt, and traditional prompt discussed in section 4.3 to the dataset by Fu and Ollagnier (2025). The original prompt templates are kept. While Hua et al. (2024) mention that they used examples from outside the test split to develop few-shot classifiers for each summary type, the examples are not given in the published prompt templates, and as these examples are specific to their dataset, we do not consider adding examples. Similar to the settings of our method for summary generation, we prompt Qwen to generate summaries and obtain the final results with “all turns”. The results are shown in Table 5. It is clear that their best performing method, procedural prompt, achieves the highest performance among the three. In comparison, with our proposed methods, in the case of msg_prag_full, the performance with CD is 6.47% higher and the performance with select+synthesize is 4.37% higher in F1 score than their best performing method. From the table, it can be seen that with our proposed methods for summarization (the last two rows), the performance is not high when summary generation is based on the messages alone, and pragmatic information is vital for our higher performance.

5.3.2 Results on the Dataset by Hua et al. (2024)

Since Hua et al. (2024) focus on results at the last turn only, we apply their best performing method on their dataset to obtain results for the

Summary Type		P	R	F1
procedural prompt	last turn only	63.57 (1.09)	62.00 (1.15)	60.86 (1.36)
	all turns	58.85 (0.52)	58.77 (0.52)	58.53 (0.53)
CD	last turn only	68.02 (3.40)	62.75 (2.22)	59.84 (2.54)
	all turns	60.59 (0.96)	60.46 (0.95)	60.43 (0.96)

Table 6: Results on the dataset by Hua et al. (2024). Standard deviations across the four summaries are presented in parentheses.

dynamic paradigm of conversation derailment prediction. Moreover, we evaluate our methods on their dataset⁵. Based on Table 3, we choose our CD method for this set of experiments. Similar to Hua et al. (2024), GPT-3 (*gpt-3.5-turbo*) is used for summary generation. Four summaries are generated for each turn dynamically and the classifier’s performance is averaged over the four summaries. The results are shown in Table 6. The metric used by Hua et al. (2024) is accuracy, which is 67.3 (± 2.63). We failed to reproduce the results and obtained an accuracy score of 62.00 (± 1.15) with their procedural prompt method. It is worth noting that we did not incorporate examples into the prompt template.

As is clear from the table, results obtained with “all turns” are generally lower than with “last turn only”, which can be explained by the fact that the first few exchanges provide weaker signals than later messages for conversation derailment prediction (Kementchedjhieva and Søgaard, 2021). With our method, the standard deviation is higher in the case of “last turn only”, which might be related to additional use of pragmatic information. Moreover, the results obtained with our method are slightly lower in terms of “last turn only” metric (but showing large standard deviation), and the results are higher ($\approx 2\%$ higher) in case of “all turns” than the best performing method by Hua et al. (2024).

6 Conclusions and Future Work

We propose a systematic approach to capturing conversation dynamics for the task of conversation derailment prediction, built on the framework of annotating pragmatic information of conversations proposed by Fu and Ollagnier (2025), which is

⁵The annotations of pragmatic information generated automatically at intermediate steps are available at: <https://github.com/yingxueF/SummConvDynamicsForecastingConvDerailment>

grounded in linguistic theories. Experimental results demonstrate the effectiveness of our approach.

In future work, a question of practical and theoretical significance is how to identify strategies to react proactively in face of prompts of conversation derailment risks.

Limitations

With our method, it is necessary to perform annotation of pragmatic information first, which is prone to errors with automatic means. The prompt templates could be lengthy and API calls are costly if a conversation is long. It might be useful to configure a buffer to save conversation trajectory incrementally. However, given the dynamic nature of conversations, the interpretation of previous messages may shift as new messages arrive, so the conversation history cannot be stored statically.

Moreover, the ChangeMyView corpus, while valuable for its protocol and established baselines, represents a particular genre of online discourse. Future work should try to incorporate varied styles of conversations and test proposed approach on diverse platforms and conversation types, such as X (formerly Twitter) threads, Facebook comments, and Discord conversations. We hypothesize that pragmatic signals may manifest differently across contexts, for example, the role of politeness strategies may vary between formal debates and casual chats.

Additionally, relying on proprietary models for pragmatic annotation raises concerns about annotation bias, reproducibility, and model dependence. As noted in footnote 1, we initially experimented with *Qwen2.5-1.5B-Instruct* (the latest release at the time) for pragmatic annotation, but found its performance unacceptable for this task: the model consistently omitted messages when processing multi-turn conversations, likely due to limitations in long-context handling. This motivated our choice of using GPT-4 for the more demanding pragmatic annotation task while employing Qwen for summary generation. To improve transparency, the intermediate annotations of pragmatic information have been released.

Furthermore, the paper can be enhanced by showing the performance of an end-to-end system. However, we separate the step of annotating pragmatic information from summary generation and conversation derailment prediction to better control the intermediate results.

Ethical Considerations

Privacy and data ethics are critical in deployment of systems for conversation derailment prediction and ASB detection systems in general. Even with anonymization, re-identification risks remain, particularly for marginalized or vulnerable groups. Risks include user profiling, where models infer the likelihood of individuals engaging in or becoming targets of ASBs based on prior behavior, linguistic cues, and engagement patterns. While such approaches may support early intervention or moderation, they also introduce significant ethical risks. To mitigate these risks, ASB prediction systems must adhere to core principles of fairness, transparency, and accountability, as emphasized in regulatory frameworks such as the GDPR (Council, 2016), and should align with ongoing efforts in explainable AI (Parasar et al., 2024) to ensure that model decisions are interpretable and contestable (Ollagnier, 2026).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Enas Altarawneh, Ameeta Agrawal, Michael Jenkin, and Manos Papagelis. 2023. [Conversation derailment forecasting with graph convolutional networks](#). In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 160–169, Toronto, Canada. Association for Computational Linguistics.
- John Langshaw Austin. 1962. *How to do things with words*. Oxford University Press.
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.
- Betty J. Birner. 2025. Pragmatics. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.
- Bharathi Raja Chakravarthi, Prasanna Kumareshan, Ruba Priyadarshini, Paul Buitelaar, Asha Hegde, Hosahalli Shashirekha, Saranya Rajiakodi, Miguel Ángel García, Salud María Jiménez-Zafra, José García-Díaz, Rafael Valencia-García, Kishore Ponnusamy, Poorvi Shetty, and Daniel García-Baena. 2024. [Overview of third shared task on homophobia and transphobia detection in social media comments](#). In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 124–132, St. Julian’s, Malta. Association for Computational Linguistics.
- Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. [Trouble on the horizon: Forecasting the derailment of online conversations as they develop](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4743–4754, Hong Kong, China. Association for Computational Linguistics.
- Jonathan P Chang, Charlotte Schluger, and Cristian Danescu-Niculescu-Mizil. 2022. Thread with caution: Proactively helping users assess and deescalate tension in their online discussions. *Proceedings of the ACM on human-computer interaction*, 6(CSCW2):1–37.
- Xi Chen and Shuo Wang. 2025. [Pragmatic inference chain \(PIC\) improving LLMs’ reasoning of authentic implicit toxic language](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5837–5852, Suzhou, China. Association for Computational Linguistics.
- European Council. 2016. Eu regulation 2016/679 general data protection regulation (gdpr). Official Journal of the European Union. Regulation (EU) 2016/679.
- Thomas Davidson, Dana Warmesley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.
- Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. [COLD: A benchmark for Chinese offensive language detection](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. [Latent hatred: A benchmark for understanding implicit hate speech](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Paula Fortuna, Juan Soler, and Leo Wanner. 2020. [Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.
- Yingxue Fu and Anaïs Ollagnier. 2025. Contextualizing toxicity: An annotation framework for unveiling

- pragmatics in conversations of online discussion forums.
- Lei Gao and Ruihong Huang. 2017. [Detecting online hate speech using context aware models](#). In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.
- Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.
- Yilun Hua, Nicholas Chernogor, Yuzhe Gu, Seoyeon Jeong, Miranda Luo, and Cristian Danescu-Niculescu-Mizil. 2024. [How did we get here? summarizing conversation dynamics](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7452–7477, Mexico City, Mexico. Association for Computational Linguistics.
- 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 the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2818–2823, Brussels, Belgium. Association for Computational Linguistics.
- Yiping Jin, Leo Wanner, and Aneesh Moideen Koya. 2025. [What the#?!: Disentangling hate across target identities](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 199–221.
- David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. [A just and comprehensive strategy for using NLP to address online abuse](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics.
- Yova Kementchedjheva and Anders Søggaard. 2021. [Dynamic forecasting of conversation derailment](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Huije Lee, Young Ju Na, Hoyun Song, Jisu Shin, and Jong Park. 2022. [ELF22: A context-based counter trolling dataset to combat Internet trolls](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3530–3541, Marseille, France. European Language Resources Association.
- Huije Lee, Hoyun Song, Jisu Shin, Sukmin Cho, SeungYoon Han, and Jong C. Park. 2024. [Towards effective counter-responses: Aligning human preferences with strategies to combat online trolling](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11670–11686, Miami, Florida, USA. Association for Computational Linguistics.
- Stefano Menini, Giovanni Moretti, Michele Corazza, Elena Cabrio, Sara Tonelli, and Serena Villata. 2019. [A system to monitor cyberbullying based on message classification and social network analysis](#). In *Proceedings of the Third Workshop on Abusive Language Online*, pages 105–110, Florence, Italy. Association for Computational Linguistics.
- Lin Miao, Mark Last, and Marina Litvak. 2020. [Detecting troll tweets in a bilingual corpus](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6247–6254, Marseille, France. European Language Resources Association.
- Luis Gerardo Mojica de la Vega and Vincent Ng. 2018. [Modeling trolling in social media conversations](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Debora Nozza and Dirk Hovy. 2023. [The state of profanity obfuscation in natural language processing scientific publications](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3897–3909, Toronto, Canada. Association for Computational Linguistics.
- Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. [An in-depth analysis of implicit and subtle hate speech messages](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.
- Anaïs Ollagnier. 2026. [Antisocial behavior prediction: A survey and practical guide](#). In *The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment Social Media Analysis (WASSA 2026)*, pages 235–251, Rabat, Morocco. Association for Computational Linguistics.
- Anaïs Ollagnier, Elena Cabrio, Serena Villata, and Catherine Blaya. 2022. [CyberAgressionAdo-v1: a dataset of annotated online aggressions in French collected through a role-playing game](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 867–875, Marseille, France. European Language Resources Association.
- OpenAI. 2025. [GPT-5 system card](#). Technical report, OpenAI. Accessed: October 1, 2025.
- Ashish Parasara, Vandana Sharma, and Preety Shoran. 2024. [Analyzing hate speech detection using explainable ai](#). In *2024 IEEE 11th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, pages 1–5.

- John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. [Deeper attention to abusive user content moderation](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135, Copenhagen, Denmark. Association for Computational Linguistics.
- John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. [Toxicity detection: Does context really matter?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.
- Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.
- Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. [A benchmark dataset for learning to intervene in online hate speech](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. [Social bias frames: Reasoning about social and power implications of language](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.
- Anna Schmidt and Michael Wiegand. 2017. [A survey on hate speech detection using natural language processing](#). In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.
- John R. Searle. 1975. A taxonomy of illocutionary acts. In K. Gunderson, editor, *Language, Mind and Knowledge*, pages 344–369. University of Minnesota Press.
- Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. [Creating a WhatsApp dataset to study pre-teen cyberbullying](#). In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59, Brussels, Belgium. Association for Computational Linguistics.
- Neha Srikanth, Rupak Sarkar, Heran Mane, Elizabeth Aparicio, Quynh Nguyen, Rachel Rudinger, and Jordan Boyd-Graber. 2024. [Pregnant questions: The importance of pragmatic awareness in maternal health question answering](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7253–7268, Mexico City, Mexico. Association for Computational Linguistics.
- Kanishk Verma, Tijana Milosevic, Keith Cortis, and Brian Davis. 2022. [Benchmarking language models for cyberbullying identification and classification from social-media texts](#). In *Proceedings of the First Workshop on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference*, pages 26–31, Marseille, France. European Language Resources Association.
- Zeerak Waseem, Thomas Davidson, Dana Warmusley, and Ingmar Weber. 2017. [Understanding abuse: A typology of abusive language detection subtasks](#). In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.
- Zeerak Waseem and Dirk Hovy. 2016. [Hateful symbols or hateful people? predictive features for hate speech detection on Twitter](#). In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.
- Deirdre Wilson and Dan Sperber. 2004. Relevance theory. In Laurence R. Horn and Gregory Ward, editors, *The Handbook of Pragmatics*, pages 607–632. Blackwell Publishing, Oxford.
- Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.
- Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. [HARE: Explainable hate speech detection with step-by-step reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505, Singapore. Association for Computational Linguistics.
- Michael Yoder, Lynnette Ng, David West Brown, and Kathleen Carley. 2022. [How hate speech varies by target identity: A computational analysis](#). In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 27–39, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Jiaqing Yuan and Munindar P. Singh. 2023. [Conversation modeling to predict derailment](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 17:926–935.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 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)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

Yunfan Zhang, Kathleen McKeown, and Smaranda Muresan. 2025. [Forecasting conversation derailments through generation](#). In *Proceedings of the 18th International Natural Language Generation Conference*, pages 699–715, Hanoi, Vietnam. Association for Computational Linguistics.

A Appendix: Prompt Template for Annotating Pragmatic Information

You are a language expert.

a) Given a thread of messages from an online forum, generate 2-3 or more, if possible, most important pragmatic inferences for EACH message.

The inferences can be sentiments, such as emotional state or stance towards some issues, and hidden knowledge from the message, including stereotypes and social constructs, pre-assumptions held by the author of the message, and implications derivable from the message. Try to avoid trivial inferences.

Be CONCISE. For example, given a sentence “What kind of music should I play to my baby in the womb?”, you need to generate inferences like the following:

1. Babies can hear sound in the womb.
2. Babies can differentiate different sounds.
3. Hearing music positively influences fetal development.
4. Certain kinds of music are more beneficial to babies in the womb than the others.

The following inferences are also possible. However, these are trivial inferences that do not contribute to conveying the main point.

- There is a baby in the interlocutor’s womb.
- Different kinds of music are available.
- Music is something that can be played.

b) When giving an inference, you need to determine the “type” of the inference from one of the following:

1. representatives: statements that tell how things are, focusing on the truth of a proposition.
2. directives: statements attempting to persuade the hearer to do something, such as ordering, advising, and warning the hearer.
3. commissives: statements that commit the speaker to a course of action, such as actions of vowing, promising, offering to do something, or negatively, refusing or preventing someone from doing something by threatening.
4. expressives: statements expressing a psychological state, such as apologizing, congratulating, condoling.
5. declarations: statements effecting an immediate change in the institutional state of affairs (firing from employment, christening). e.g. “I declare you husband and wife”, and “You are fired”!

c) If no pragmatic inferences are considered important for the conversation but only literal interpretation of the messages is used by the conversation participants, simply output “literal” and “none” in “type”.

d) Determine the most salient inference for a message BASED ON its following message, and copy its type to an entry “inference_type”.

e) Determine whether the most salient inference is agreed in the reply message. Put “yes” or “no” or “uncertain” (when no following messages) in an entry “as_intended”.

f) Determine whether the most salient inference belongs to a “presupposition” (PRE, something taken to be true by the speaker, without proving) or “implicature” (IMP, something the speaker tries to convey with the message).

Your output should be in a valid json format:

```

{"turn_id_1": "Pragmatic_Inferences": {}},
{"turn_id_2": {...}}, specifically:
{"turn_id_1": {"Pragmatic_Inferences": { "1": {
"content": "...", "type": "..."},
"2": { "content": "...", "type": "..."},
"3": {"content": "...", "type": "..."}},
"as_intended": "yes",
"aggressive": "CAG",
"most_salient_inference": "1",
"inference_type": "...",
"PRE/IMP": "IMP"},
"turn_id_2": {"Pragmatic_Inferences": { "1": {
"content": "...", "type": "..."},
"2": { "content": "...", "type": "..."} }},

```

```
"as_intended": "yes",
"aggressive": "CAG",
"most_salient_inference": "1",
"inference_type": "...",
"PRE/IMP": "IMP" } } }
```

Here is a thread of messages from an online forum:

B Appendix: An Example of Annotations of Pragmatic Information

(A) *"post_title": Why are German state officials tend to behave friendly to uneducated/unlawful immigrants while being hostile against people who visit the country for business purposes?...*

"post_content": I am not saying that they should behave hostile against "the immigrants" but I wonder why they are so hostile against the second group?

Annotated pragmatic inferences:

"1": "content": "The author perceives a systemic bias among German state officials favoring uneducated or unlawful immigrants over business visitors.", "type": "representatives",

"2": "content": "The author assumes that business visitors should be treated at least as well as, if not better than, uneducated or unlawful immigrants.", "type": "representatives",

"3": "content": "The author is dissatisfied or feels unfairly treated by German state officials during their visit.", "type": "expressives"

```
most_salient_inference: 1
inference_type: representatives
as_intended: no
PRE/IMP: IMP
toxicity: no
```

(B) *Who are they?*

How often did it happen?

How many people where involved?

Are you sure this wasn't a singular incident?

Did you ask yourself if you might have done something wrong?

There are a fuck ton of reasons for this.

Racism is probably the last one...

Annotated pragmatic inferences:

"1": "content": "The responder doubts the generalization made by the original poster and questions the validity and frequency of the described behavior.", "type": "representatives",

"2": "content": "The responder suggests that the original poster may be misattributing the cause

of the incident, downplaying racism as a likely explanation.", "type": "representatives",

"3": "content": "The responder implies that the original poster may have contributed to the negative experience.", "type": "representatives"

```
most_salient_inference: 1
inference_type: representatives
as_intended: yes
PRE/IMP: IMP
toxicity: yes
```

C Appendix: Template for Integrating Pragmatic Information with Messages

When extracting pragmatic information from annotations, we focus on the content of the inference (`{pragmatic_inf_content_txt}`) and its illocutionary act type (`{pragmatic_inf_typ}`).

When no pragmatic inferences are generated, and a label "literal" is given, the extracted pragmatic information is: `f"The comment does not contain any pragmatic inferences; it is literal." f"The speaker said: {content} "`

When only one pragmatic inference is generated, the pragmatic inference content (`{pragmatic_inf_content_txt}`) is: `"f"A pragmatic inference can be made. It is of {pragmatic_inf_typ} speech act type." f"The inferred meaning is: {pragmatic_inf_content} "`

When more than one pragmatic inference is annotated, the pragmatic inference content for the message is: `(f"A pragmatic inference can be made. It is of {pragmatic_inf_typ} speech act type." f"The inferred meaning is: {pragmatic_inf_content} ")`

`f"Another pragmatic inference can be made. It is of {pragmatic_inf_typ} speech act type. " f"The inferred meaning is: {pragmatic_inf_content} ")`

`f"{pragmatic_inf_content_txt} ": set in the above manner`

`f"{most_salient_inf_txt} ": (f"For the conversation flow, the most salient pragmatic inference is {most_salient_inference_text} " f"The predominant inference type is {inf_typ} speech act type.")`

`f"{as_intended_inf_txt} ": "The speaker's most salient message is (not) agreed by the listener."/ "It is uncertain whether the speaker's most salient message is agreed by the listener."`

`f"{PRE_IMP_txt} ": "The most salient inference is an implicature/a presupposition. "`

C.1 Extracting Messages Only (`msg_only`)

`f"This is message {idx}. "`

f"The message is authored by {author}. "
f"The content of the message is: {content} "

C.2 Extracting Full Pragmatic Information without Using Messages (prag_only)

f"This is message {idx}. "
f"The message is authored by {author}. "
f"{pragmatic_inf_content_txt} "
f"{most_salient_inf_txt} "
f"{PRE_IMP_txt} "
f"{as_intended_inf_txt} "

C.3 Integrating Full Pragmatic Information with Messages (msg_prag_full)

f"This is message {idx}. "
f"The message is authored by {author}. "
f"The content of the message is: {content} "
f"{pragmatic_inf_content_txt} "
f"{most_salient_inf_txt} "
f"{PRE_IMP_txt} "
f"{as_intended_inf_txt} "

C.4 Only Keeping Most Salient Pragmatic Inference (most_salient_only)

f"This is message {idx}. "
f"The message is authored by {author}. "
f"{most_salient_inference_text} "
f"{PRE_IMP_txt} "
f"{as_intended_inf_txt} "

C.5 Omitting PRE/IMP and as_intended Annotations (msg_prag_salient)

f"This is message {idx}. "
f"The message is authored by {author}. "
f"The content of the message is: {content} "
f"{pragmatic_inf_content_txt} "
f"{most_salient_inf_txt} "

C.6 Integrating Most Salient Inference with Messages (msg_salient)

f"This is message {idx}. "
f"The message is authored by {author}. "
f"The content of the message is: {content} "
f"{most_salient_inf_txt} "

C.7 Integrating Pragmatic Inferences with Messages (msg_prag), Omitting Most Salient Inference

f"This is message {idx}. "
f"The message is authored by {author}. "
f"The content of the message is: {content} "
f"{pragmatic_inf_content_txt} "

D Appendix: Prompt Template for Generating Summaries with Select+Synthesize

Task 1: Select messages that are most relevant for understanding the trajectory of the conversation. An informative summary must not simply identify these aspects separately, but should also describe how they evolve and interrelate throughout a conversation. You must SYNTHESIZE different aspects across utterances and participants. The final purpose of the summary is to predict the possible aggressiveness of the conversation flow.

Task 2: Based on the summary for the conversation so far, determine if the following conversation will finally turn toxic (TRUE) or non-toxic (FALSE) finally. You MUST output your answer in the following exact format:

Summary: <summary text>

Answer: <TRUE or FALSE>

RULES:-The "Answer:" line MUST contain EXACTLY either "TRUE" or "FALSE".

- Do NOT add quotes, explanations punctuation, or any other text after TRUE or FALSE.

- Do NOT include any text outside the two required lines.

- If you are unsure, choose the best guess but still output only TRUE or FALSE.'

E Appendix: Prompt Template for Generating Summaries with CD

Task 1: Select messages that are most relevant for understanding the trajectory of the conversation. Focus on tones in messages, such as "polite", "rude", "aggressive", "condescending", or "sarcastic", tonal changes, such as "tension maintained", "escalated", "remained argumentative", patterns of interaction that are indicative of future trajectories, such as "had a brief exchange" or "entered into a lengthy discussion", conversation strategies, such as posing a question, questioning the other's logic, or expressing uncertainty, and so on. An informative summary must not simply identify these aspects separately, but should also describe how they evolve and interrelate throughout a conversation. You must SYNTHESIZE different aspects across utterances and participants. The final purpose of the summary is to predict the possible aggressiveness of the conversation flow.

Task 2: same as Appendix D

F Appendix: Prompt Templates Used in the Study by Hua et al. (2024)

Procedural prompt

Task 1: Write a short summary capturing the trajectory of an online conversation.

Do not include specific topics, claims, or arguments from the conversation. The style you should avoid:

Example Sentence 1: “Speaker1, who is Asian, defended Asians and pointed out that a study found that whites, Hispanics, and blacks were accepted into universities in that order, with Asians being accepted the least. Speaker2 acknowledged that Asians have high household income, but argued that this could be a plausible explanation for the study’s findings. Speaker1 disagreed and stated that the study did not take wealth into consideration.” This style mentions specific claims and topics, which are not needed.

Instead, do include indicators of sentiments (e.g., sarcasm, passive-aggressive, polite, frustration, attack, blame), individual intentions (e.g., agreement, disagreement, persistent-agreement, persistent-disagreement, rebuttal, defense, concession, confusion, clarification, neutral, accusation) and conversational strategies (if any) such as ‘rhetorical questions’, ‘straw man fallacy’, ‘identify fallacies’, and ‘appealing to emotions.’ The following sentences demonstrate the style you should follow:

Example Sentence 2: “Both speakers have differing opinions and appeared defensive. Speaker1 attacks Speaker2 by diminishing the importance of his argument and Speaker2 blames Speaker1 for using profane words. Both speakers accuse each other of being overly judgemental of their personal qualities rather than arguments.”

Example Sentence 3: “The two speakers refuted each other with back and forth accusations. Throughout the conversation, they kept harshly fault-finding with overly critical viewpoints, creating an intense and inefficient discussion.”

Example Sentence 4: “Speaker1 attacks Speaker2 by questioning the relevance of his premise and Speaker2 blames Speaker1 for using profane words. Both speakers accuse each other of being overly judgemental of their personal qualities rather than arguments.”

Overall, the trajectory summary should capture the key moments where the tension of the conversation notably changes. Here is an example of a

complete trajectory summary.

Trajectory Summary:

Multiple users discuss minimum wage. Four speakers express their different points of view subsequently, building off of each other’s arguments. Speaker1 disagrees with a specific point from Speaker2’s argument, triggering Speaker2 to contradict Speaker1 in response. Then, Speaker3 jumps into the conversation to support Speaker1’s argument, which leads Speaker2 to adamantly defend their argument. Speaker2 then quotes a deleted comment, giving an extensive counterargument. The overall tone remains civil.

Now, provide the trajectory summary for the conversation.

Zero-shot prompt

Task 1: Now, summarize this conversation. Remember, do not include specific topics, claims, or arguments from the conversation. Instead, try to capture the speakers’ sentiments, intentions, and conversational/persuasive strategies. Limit the trajectory summary to 80 words.

Traditional prompt

Task 1: Summarize this transcript of online conversation in 80 words.

G Appendix: An Example Showing the Effect of Incorporating Pragmatic Information

We provide an example below to show qualitatively the effect of incorporating pragmatic information in the summary.

There are four messages in a conversation that finally turns toxic. The model will make a prediction about the risk of the whole conversation going toxic. In this example, the model makes a prediction when the second message appears.

For the setting `msg_prag` (using `select+synthesize` summarization method):

→ Processing comment 2/4

[Previous message 1]

This is message 1. The message is authored by speaker1. The content of the message is: “post_title”: I figured out what my feelings should be about that guy who decided to go into my bed.

“post_content”: The verdict is I don’t know. I was unconscious the whole time so I can’t know. I can’t say that he did assault me and I can’t say he didn’t assault me. I don’t have any real memories from when I was unconscious. **It’s just an I don’t know. **

A pragmatic inference can be made. It is of expressives speech act type. The inferred meaning is: The speaker is experiencing uncertainty and confusion about how to feel regarding a potentially traumatic event. Another pragmatic inference can be made. It is of representatives speech act type. The inferred meaning is: The speaker believes that without memory or evidence, it is impossible to make a definitive judgment about whether an assault occurred. Another pragmatic inference can be made. It is of representatives speech act type. The inferred meaning is: The speaker implicitly rejects external pressure to label the event as assault or not assault without personal certainty.

[Current message]

This is message 2. The message is authored by speaker2. The content of the message is: I know. He's a rapist. A pragmatic inference can be made. It is of representatives speech act type. The inferred meaning is: The speaker is making a strong moral judgment, labeling the man as a rapist based on the described situation. Another pragmatic inference can be made. It is of representatives speech act type. The inferred meaning is: The speaker assumes that entering someone's bed while they are unconscious constitutes rape, regardless of memory or evidence.

Generated Summary: In this conversation, the first participant expresses uncertainty and confusion about a traumatic event involving a potential sexual assault, using pragmatic inferences such as expressing doubt and rejecting external judgments. The second participant then makes a strongly negative moral judgment, assuming the event qualifies as rape due to the circumstances described. This shift towards a more aggressive stance suggests a potential escalation in the conversation.

gold: toxic | predicted: toxic

Summary (without pragmatic information): In this conversation, both parties are expressing strong opinions about the same incident involving sexual assault. The first participant expresses doubt and uncertainty ("I don't know"), while the second participant makes a clear moral judgment ("He's a rapist"). There is no indication of escalating hostility or aggression; instead, there seems to be a back-and-forth exchange of views. Both speakers seem equally committed to their positions.

gold: toxic | predicted: non-toxic