Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Consequences of Conflicts in Online Conversations

Kristen M. Altenburger[1], Robert E. Kraut[2], Shirley Anugrah Hayati[3], Jane Dwivedi-Yu[1], Kaiyan Peng[1], Yi-Chia Wang[1]

[1]Meta
[2]Carnegie Mellon University
[3]University of Minnesota

Abstract

Interpersonal conflicts occur frequently in both offline and online groups, with conditions for conflict especially ripe online. This research attempts to understand the consequences of online group conflict and reporting it to group administrators, both for the protagonists in the conflict and observers. If group conflict is aversive, then group members should reduce their group participation after observing conflict. Theories of imitation and behavioral mimicry suggest that even onlookers will exhibit more conflict and negative language after observing conflict conversations in their group. In contrast, theories of deterrence suggest that both the instigator of the conflict and onlookers will reduce their conflict and onlookers might even increase their engagement if conflicts are reported to group administrators. The current study uses de-identified and aggregated data from Facebook group conversations and Mahalanobis distance matching to test these ideas. Results are consistent with the hypothesis that conflict in group conversations reduces engagement within the group and increases the amount of conflict and the negativity of language users express in the group. However, inconsistent with deterrence theories, conflict and language negativity increase and group engagement decreases when conflict is reported to group administrators.

Introduction

Research in organizational behavior distinguishes between task conflict, in which individuals in a group disagree on which goals to pursue and how to achieve them, and relationship conflict, which involves interpersonal hostility between members of a group. Conflict is surprisingly common in both offline and online groups. Although it can result from conflicts of interests or discussing contentious topics (De Dreu 2010), even disagreements on non-contentious topics can easily spiral into interpersonal conflict (Levy et al. 2022). When discussing conflict in the online context, we do not mean simple disagreement over opinions or interests, but rather more hurtful or demeaning attacks of one person on another, often through the use of personal insult and profanity. This is analogous to "relationship conflict" as used in the literature on organizational behavior.

Studies of off-line teams consistently show that relationship conflict is associated with negative group outcomes, including worse member satisfaction, worse group cohesion, and poorer productivity (De Dreu and Weingart 2003). Because relationship conflict can be unpleasant to the people involved and can lead to negative downstream consequences for the group as a whole, both off-line and online groups try to reduce it. Online groups have many characteristics that make conflict and interpersonal attacks especially likely to emerge, such as the relative anonymity of members resulting from pseudonyms and large size, rapid turnover in membership, lack of vetting of new members, often extreme diversity in membership, and impoverished channels of communication (de la Vega and Ng 2018; Ma et al. 2019). Once conflict starts to emerge between a pair of people, the conflict might escalate in the pair and may spread to other people who were not initially involved (De Dreu 2010).

If people in online conversations and groups find interpersonal conflict unpleasant, then group managers and the owners of the platforms where these conflicts take place have an interest in reducing it; a number of online platforms have built tools or put policies in place to reduce conflict. For example, in the online context, machine learning models can identify personal attacks, insults, profanity, hate speech, and other signs of interpersonal conflict with reasonable accuracy based on both content of a conversation and its structure (Sood, Churchill, and Antin 2012; Zhang et al. 2018; Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015) and human moderators can warn (Yildirim et al. 2023) or ban (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015) group members who consistently exhibit antisocial behavior. The interface of Facebook groups allows group members to report episodes of relationship conflict to group administrators, and the Facebook platform additionally hosts a tool intended to predict conflict and automatically alert administrators (Perez 2021). Because content that triggers stronger physiological responses tends to be shared more on social media (Nelson-Field, Riebe, and Newstead 2013), online conflict may actually attract some viewers. If online conflict is indeed engaging, then episodes of conflict might be a mixed blessing—harmful to the group as a whole by increasing negativity in the group and reducing cohesion, while at the same time attracting group members who are interested in witnessing emotion-laden content.

The goal of the current research is to examine the consequences of conversational conflict and its reports for the

different people involved, including the protagonists in the conversation who wrote the comments containing conflict and other people who participated in the conversation but were not themselves writers of conflictful messages. In this work, we formulate several hypotheses regarding the consequences of online conflict that are grounded in two relevant bodies of research: *behavioral mimicry* and *deterrence*. Research on imitation, behavioral mimicry, and emotional contagion suggests that witnessing conflictful interactions online will increase the likelihood that onlookers would subsequently engage in more conflictful interaction. On the other hand, research on theories of learning in psychology and specific deterrence in criminology suggests that when conflict is reported to a group administrator the reported individual should subsequently engage in less conflict, and social learning theory and generalized deterrence suggests that onlookers would also be less likely to engage in subsequent conflict. However, the impact of reports depends upon how administrators respond to the original conflict, where responses can range from doing nothing to chastising the protagonists privately to banning them from the group for some period.

Our contributions include an analysis of the ramifications of two types of events: (1) online conflict occurring in a group and (2) reporting of online conflict within the group through a Facebook groups tool. To study the effects of these two events, we collect data prior to and following each event and use Mahalanobis distance matching (MDM) (King and Nielsen 2019; Stuart 2010) to compare matched samples of people exposed to the event with others who were not exposed. We conduct regression analyses to examine differences in language use and behavioral engagement between the matched groups. Unlike much of the previous research on online conflict that has focused on the causes of conflict (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015; Levy et al. 2022), our research focuses on its downstream consequence for those involved in the conflict and the group as a whole.

Related Work and Hypotheses

In this section, we discuss related work on measuring the outcomes and presence of conflict through linguistic features. Secondly, we review two bodies of relevant research—*imitation* and *learning theory*—that help formulate hypotheses for the outcomes of participating in and observing online conflict.

Linguistic features. In order to characterize the effects of experiencing conflict and reporting, we compare group members who are exposed to and not exposed to these events on several linguistic measures: conflict, politeness, sentiment, and anger. First, we measure the degree of conflict expressed in subsequent conversations. Second, we measure politeness, a language feature strongly related to status and the power dynamics in social and online interactions (Danescu-Niculescu-Mizil et al. 2013a; Obeng 1997; Chilton 1990; Andersson and Pearson 1999). Politeness can be a meaningful indicator of whether an interaction is going well or not (Kasper et al. 1993) and can be used as a strategic means of avoiding conflict (Yeomans et al. 2020). Third, we examine affective properties such as the degree of positive and negative sentiment expressed in the conversation. Sentiment on social media varies with both online and off-line conflict (Lucić, Katalinić, and Dokman 2020; Chambers et al. 2015; Abedin, Jafarzadeh, and Akhlaghpour 2018), and sentiment analysis has been used to quantify the degree of conflict, such as in legislative speeches (Proksch et al. 2019). Fourth, because sentiment is a gross approximation of individuals' affective state, we also analyze the amount of anger expressed in a conversation to more precisely characterize one of the specific emotions that might be triggered by conflict. Anger generally increases as conversations become more conflictful (Levy et al. 2022). We chose anger rather than other negative emotions like sadness or fear because anger can lead to escalations of interpersonal conflict by leading people try to safeguard their beliefs, make more risky choices, reject compromise, emphasize punishment or retaliation, and reduce trust (Lerner and Keltner 2001; Bodtker and Jameson 2001; Allred et al. 1997; Sharma et al. 2020).

Imitation and Conformity. Several related processes lead to the spread of behavior from one person to another, ranging from relatively unconscious emotional contagion and behavioral mimicry to more deliberate social learning and imitation. Studies have identified imitation effects in a variety of different contexts, ranging from the mimicry of non-verbal behavior and language (Chartrand and Bargh 1999; Asch 1956) to copycat suicide (Phillips 1974). Social learning theory, which argues that people acquire and perform actions through their associations with other people who are performing them, contributes an related perspective on mechanisms involved in the adoption of deviant and conforming behavior, ranging from adolescent marijuana usage (Akers and Cochran 1985) to alcohol use among the elderly (Akers and La Greca 1991).

Imitation has also been documented online (Romero, Meeder, and Kleinberg 2011; Zhu, Kraut, and Kittur 2012). Bakshy et al. (2012) found that Facebook users were significantly more likely to share a link if their friends had shared it. Imitation and conformity occur among Wikipedia editors too, including the spreading of personal attacks (Wulczyn, Thain, and Dixon 2017).

We hypothesize that these imitation effects will occur when people witness conversational conflict in online groups and will generalize to related behaviors.

- **H1**: The amount of conflict in an online group will increase following the occurrence of an episode of conversational conflict.
- **H2**: Messages exchanged in an online group will become less polite, more negative, and display more anger following the occurrence of an episode of conversational conflict.

Furthermore, negative emotions can lead to less engagement on social media (Kujur and Singh 2018) so if H1 and H2 are true, we should also observe a decrease in engagement after conflict.

- **H3**: The participation of the members in the group would

drop after they are involved in an episode of conversational conflict.

Feedback. While theories of emotional contagion and imitation focus primarily on the ways behaviors spread by observing others' behaviors, theories that incorporate feedback, with roots in research on learning (Kluger and DeNisi 1996), social learning (Bandura and Walters 1977), and criminology (Gibbs 1985), focuses on the impact of observing the consequences of actions. It is used here to reason about the likely consequence of having episodes of conflict reported to an authority figure, such as a group administrator. These feedback theories distinguish behavioral change that results from the consequences of one's own actions and observing the consequences of others actions. For example, theories of deterrence distinguish between specific deterrence, which is the impact of punishments on individuals who receive them, and general deterrence, which is the impact of the threat of such punishments on uninvolved observers (Stafford and Warr 1993). These feedback theories assume rational actors whose likelihood of performing antisocial actions will decrease to the extent that they judge the costs, including the probability and severity of negative consequences, are higher. In classic deterrence theory, the costs involve criminal sanctions, but more generally they can include other social costs, including the potential of shame and loss of respect in being called out for an action.

In particular, deterrence researchers have argued that individuals who have offended and been caught (specific deterrence) or have knowledge of others who have offended and been caught (general deterrence), are less likely to offend in the future. Although deterrence theory was designed to explain the effects of the certainty and severity of punishment on crime, the impact of deterrence seems to be greater for administrative offenses, like violations of rules, and infringements of informal social norms than for serious crime (Dölling et al. 2009). Deterrence effects have been demonstrated in online contexts for such behaviors as piracy, hacking, aggression, and spam (Higgins, Wilson, and Fell 2005; Maimon, Howell, and Burruss 2021; Xu, Xu, and Li 2016; Seering, Kraut, and Dabbish 2017). For example, when Twitter users who practice hate speech are warned about risks that their account may be suspended, they reduce their hate speech, at least temporarily (Yildirim et al. 2023). We hypothesize that after a conflict is reported, general and specific deterrence will impact subsequent messages in the following ways:

- **H4a**: Because of specific deterrence, when a person's conversation is reported to administrators as conflict, that person's subsequent content will contain less conflict and negative content and be more polite.
- **H4b**: In addition, the reported person will participate in the group less, posting, commenting, and reacting less.
- **H5**: Because of general deterrence and social learning, when one person's conversation is reported to administrators as conflict, subsequent communication by other participants in the reported conversation will contain less conflict and negative content and be more polite.
- **H6**: Because reporting a person for conflict leads to less conflict and negative content in the group, other people will engage in the group more.

Data and Conflict Detection Model

This research addresses two general research questions:

- **RQ1 Conflict:** What is the impact of being involved in conflict conversations on participants' subsequent behavior?
- **RQ2 Conflict Reports:** What is the impact of reporting a conflict to an authority on participants' subsequent behavior?

Answering these two research questions requires two sets of data: (1) Matched conversations that contain conflict versus similar conversations that do not and (2) Matched conflict-containing conversations that were reported to administrators versus similar conflict-containing conversations that were not reported.

To distinguish (1) conversations that contain conflict and those that do not, we use a conflict detection model, the details of which will be described below. We distinguish between reported and non-reported comments (2) through a comment reporting tool available to members of a Facebook group. When a user in a group views a comment, they can select "report comment to group admins" from a drop-down menu next to the comment and then select "Member Conflict" as the reason for reporting the comment.[1]

These procedures lead to three sets of threads for analysis: user-reported conflict, non-reported conflict, and non-conflict. To answer RQ1 on the impact of conflict, we compare reported and non-reported conflict to non-conflict threads, which is illustrated as "Matched for Conflict Analysis" in Figure 1. For RQ2 on the impact of reports, we compare non-reported conflict threads, and reported conflict threads, which is illustrated as "Matched for Reported Analysis" in Figure 1. The details of the final data samples are included in Table 1.

Conflict Detection Model. We use a BERT-based model (Devlin et al. 2019), trained to identify conflict comments and to classify conversations into conflict or non-conflict ones. BERT models achieve state-of-the-art results on a wide variety of natural language processing tasks (Devlin et al. 2019). Our BERT conflict model is trained using conflict reports as the ground truth. Specifically, the training dataset consists of 43,564 reported conflict comments and 43,564 randomly-sampled non-reported ones collected from April 2022 - June 2022. To build the model, we extract features from the target comment, the comments preceding the target, and the post that started the conversation. For each target comment, we retrieve up to 20 of its parent comments and concatenate them to the target comment. The model performance on a 20% held-out test set is highly accurate, with an F1-score of 93.41% where 50% is chance (false negative rate = 6.20%; false positive rate = 6.99%). An example of two threads that would be classified as conflict are shown

[1] See https://www.facebook.com/help/1380418588640631 for a description of the content reporting tool.

in Figure 1, Conversations 1 & 2, in which the final comments in the threads are classified as conflict by virtue of the toxic manner in which the disagreements are expressed. In Conversation 1, a disagreement about use of glue traps for killing mice ends with a comment saying " [J]ust give a fuck about it without a damn voice so the mouse won't suffer."

Data collection. In order to analyze the outcomes of conflict conversations and reporting (i.e., the "treatments"), we collect data from three different periods. the 2-week *treatment period* from June 1–14, 2022 during which conflict and reports occurred, a 4-week *pre-treatment period* from May 4–31, 2022, and a 4-week post-treatment *observation period* from June 15, 2022 – July 12, 2022 . We collect the treatment threads from the treatment period, features used for matching from the pre-treatment period, and outcome measures from the post-treatment period.

To answer RQ1, we compare group members' behaviors after they are exposed to conflict conversations with those exposed to matched non-conflict conversations. Conflict conversations are identified via the conflict detection model described in the next section. For RQ2, we compare group members' behaviors after they are exposed to a conflict conversation that is reported to the group administrator by a group member with those exposed to the matched conflict conversation which is NOT reported.

Conversations involving conflict can differ from non-conflictful conversation on many confounding factors beside the conflict they contain; for example, people engaged in conflict generally post more frequently than those who do not (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015). Similarly conversation reported for having conflict can differ from non-reported conversation on many confounding factors, including the severity of the conflict. Because our research involves observational data, without random assignment of conflict to threads or reports to conflictful threads, we use Mahalanobis Distance Matching (MDM) as a causal inference strategy to infer the impact of conflict and conflict reports by matching conversations on may potential confounds (King and Nielsen 2019; Stuart 2010; Sekhon 2009)

Broader perspective and ethical considerations. An internal research board reviewed the study's research ethics and privacy practices prior to its start. In agreeing to its terms of service, Facebook users allow the type of analysis done in this paper to better understand how people use Meta products, to improve the products, and to promote their safety, security, and integrity, including combating harmful user conduct. To preserve users' privacy, data consists of de-identified comments collected only from public groups (any size) and large (>32 members) visible, private groups (i.e., ones where anyone can see the group's name and description and request membership). When writing up this paper, to preserve privacy, we only included example comments from public groups and paraphrased them so that they could not be found through an internet search (Bruckman 2002). The analysis relies, in part, on user-reported conflict data. As discussed in the limitation section, this reported conflict label is not perfect, as users might falsely report conflict. We relied on a qualitative review of comments to assess label quality. The results of this research aims to understand the effects of conflict and reporting such conflicts in order to better prevent conflict and minimize negative consequences. We note that before putting conflict mitigation tools into production, such interventions would need to be evaluated for their ethical impact and to ensure that they are done to improve users' experience rather than to manipulate a user's perception or preferences.

	Before Matching		*After Matching*	
	#Threads	**#Groups**	**#Threads**	**#Groups**
Conflict Data				
Conflict	40,086	2,803	28,071	2,803
Non-conflict	111,069	2,803	28,071	2,803
Reported Data				
Reported conflict	5,592	3,955	2,447	1,421
Non-rep. conflict	387,042	3,955	2,447	1,421

Table 1: Statistics on the dataset for both conflict and reported analyses before and after matching.

Methodology

In this section, we describe how we apply Mahalanobis distance matching to balance confounding factors when comparing the effects of participating in conflict threads versus non-conflict ones and in reported conflict threads versus non-reported one (see the Mahalanobis Distance Matching within Groups subsection), describe the features we use for matching and in the regression analysis (see the Matching subsection), and explain our regression analysis method (see the Regression Outcomes Analysis subsection). Unlike (Levy et al. 2022) which examined *predictors* of conflict in a conversation, here we examine the *impact* of conflict.

Mahalanobis Distance Matching within Groups

Matching accounts for possible selection-biases, in which pre-existing differences among participants lead them to be exposed to the treatment or not. By using a matching methodology, we can reduce imbalance due to other confounding variables, such as the nature of the group where the conversation took place, and participants' engagement and language before they were exposed to the "treatments". To do this, we implement clustered matching, following (Arpino and Cannas 2016), by matching threads within the same Facebook group. This cluster-level matching is appropriate since it discourages matching dissimilar threads across different Facebook groups. The distance between a treated unit and a control unit is computed using Mahalanobis Distance Matching (MDM) (Mahalanobis 1936; King et al. 2011; Cochran and Rubin 1973; Rubin 1979), a modified Euclidean distance that measures how many standard deviations a point is from the mean of the group of points.

MDM is a quasi-experimental method that uses statistical techniques to construct an artificial control group by matching each treated unit with a very similar non-treated unit.

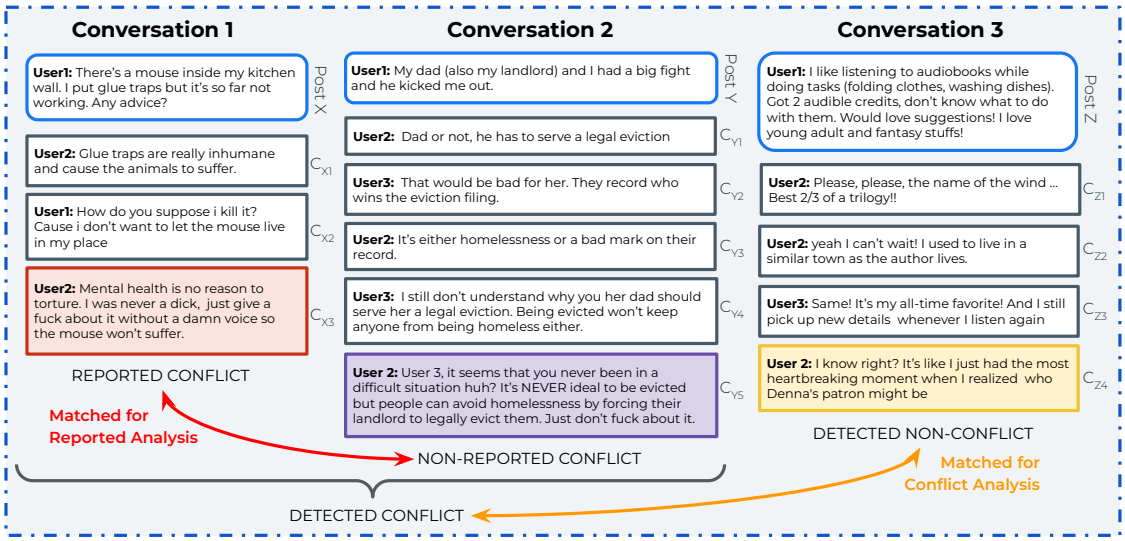

Figure 1: For the conflict analysis, we match comments detected as containing conflict by our model vs. comments detected as non-conflict and were not reported. For the reported analysis, we match the detected conflict comments that were reported vs. ones that were not reported.

This matching procedure aims to approximate a random-assignment experiment (Rosenbaum and Rubin 1983) comparing similar conflict threads to non-conflict ones and reported conflict threads to non-reported ones. After this matching procedure, we then run all analyses on the matched sample and discard unmatched observations. In this study, the *"treatment"* for RQ1 is participating in a conversation containing conflict (versus one not containing conflict) and the *"treatment"* for RQ2 is participating in a conflictful conversation that was reported to an administrator (versus one that was not reported). Because we are interested in understanding the impact of participating in different types of conversation on participants' future behavior, we implement matching at the thread-level since comment threads are the mechanisms through which conversations occur in Facebook groups. Thread-level matching has the consequence of limiting matched users to those who are part of similar conversations.

Conflict vs. non-conflict. The following steps describe how we matched threads containing conflict with similar threads without conflict:

Step 1: We applied the conflict detection model to the full data set from the treatment period. Comments with conflict probability score >0.95 are defined as conflict comments, while comments with probability score <0.45 are marked as non-conflict comments. These thresholds correspond to the top 30% and bottom 30% percentiles of the conflict distribution. We choose the top 30% percentile as a conflict threshold because the median conflict probability score for reported conflict data is 0.94. These thresholds produce 40,086 conflict threads and 111,069 non-conflict comment threads before matching (see Table 1). We exclude comments with conflict probability scores between 0.45 and 0.95 because they are generally too ambiguous to be considered as non-conflict comments.

Step 2: Then we conduct the clustered matching approach using all features explained in the Matching Section. Dropping conflict threads with no good non-conflict matches results in around 28k conflict comment threads matched with 28k non-conflict comment threads (see Table 1).

We iterate these steps to obtain the best matching settings by comparing the absolute standard mean difference (ASMD) for potential confounds before and after matching. Although there is no consensus in the literature on the value of a standardized difference that denotes important residual imbalance between treated and untreated units, some have argued a ASMD below 0.1 is acceptable (Austin 2009). The matching procedure produced a substantially more balanced data set, with all ASMDs for potential confounds falling below 0.1, except for the number of unique members whose ASMD equated 0.33 and for the percent of females in the group whose ASMD equated to 0.13. As evident from the balance plot in Figure 2, most of the ASMDs for potential confounds are much closer to 0 after matching, indicating that the matching procedure produced a substantially more balanced data set. For example, the balance plot on the left hand side of Figure 2 shows that before matching people involved in conflict threads compared to non-conflict ones differed in the politeness of their comments in the pre-treatment period but that this discrepancy is eliminated after matching.

Reported conflict vs. Non-reported conflict. To examine RQ2 on the effects of reporting, we must compare reported and non-reported threads containing a similar degree of conflict. Consequently, we add the conflict probability scores as a feature in our matching process. The following are the steps for matching reported conflict vs. non-reported conflict:

Step 1: We label both reported conflict comments and non-reported comments with conflict probability scores from our conflict detection model. Reported conflict threads

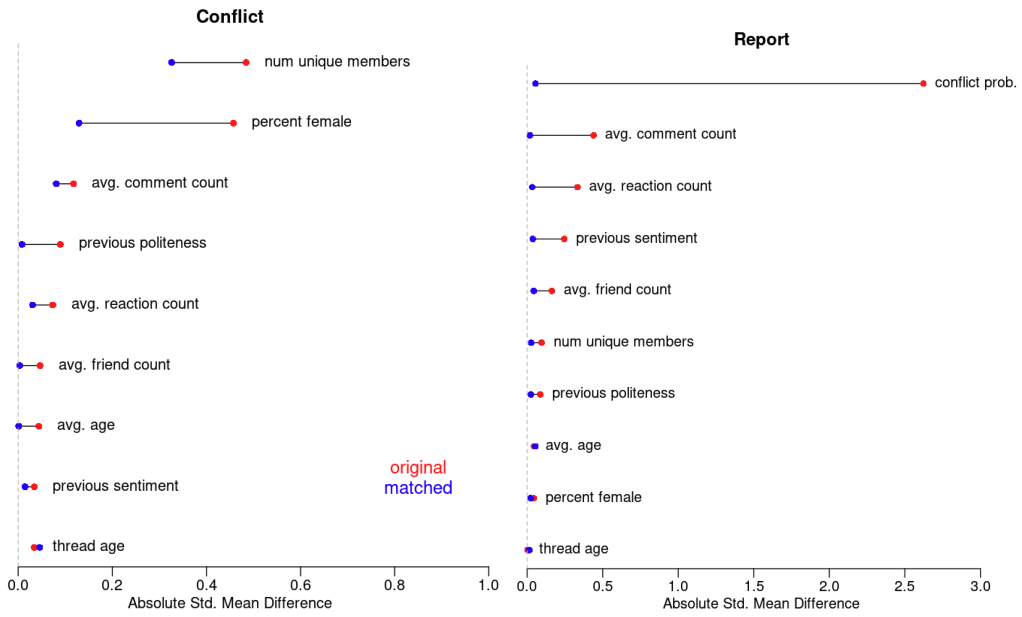

Figure 2: Balance plots for the conflict sample (left) and report sample (right) showing the covariate balance before and after Mahalanobis Distance Matching.

are our treated units, and the non-reported comments are the non-treated units.

Step 2: Then, we add the conflict probability score as one of the matching features. We set the caliper for the conflict probability score to be 0.1 so that we only collect non-reported conflict comments that have at most a 0.1 probability score difference with the reported comments.

Step 3: Finally, we conduct the clustered matching approach using the new conflict probability score and all the other features explained in the Matching Section. We also drop reported threads that cannot be matched with a similar non-reported thread, resulting in 2,447 reported conflict threads matched with 2,447 non-reported conflict threads. The matching procedure produced a substantially more balanced data set compared to the original, with all ASMDs for potential confounds, including the conflict probability score, below 0.1. We illustrate the balance plot in Figure 2 (right).

Matching

We match on four sets of features: *thread characteristics*, *thread participants' demographics*, *thread participants' activity in the group*, and *thread participants' linguistics behavior*.[2] We focus on features that are related to (1) whether a thread is conflict or not and (2) whether a conflict thread is reported or not. Note a target comment refers to the conflict comment or its matched non-conflict comment. We call users who are involved in a conversation thread "thread participants". We use average thread values for matching and the raw user-level values in the regression analysis.

Set 1: Thread characteristics Prior work examining toxic conversations on Twitter observes that they occur in deeper and larger reply trees (Saveski, Roy, and Roy 2021). Because we're studying only threads and not full reply trees, we approximate these characteristics by controlling for the number of unique participants in the thread and its age.

- *Number unique participants in the thread*. The number of unique participants in the thread, including the users who wrote the target comment, the parent comments, or the initiating post.
- *Thread age*. The difference in days between the date of the initiating post and date the target comment was created.

Set 2: Participants' demographics Prior work shows an association between user demographics and whether they are involved in a conflict thread (Levy et al. 2022). We use the average user demographics in a thread for matching and user-level demographics as control variables in the regression analyses.

- *Female percentage in the thread.* The number of self-identified female users divided by the number of unique participants in the thread.
- *Average age of thread participants.* The average age of unique participants in the thread.
- *Average friend count of thread participants.* The average number of friends of unique participants in the thread.

Set 3: Participants' activity in the group Prior work has examined participants' activity when predicting antisocial behavior (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015), controversy (Coletto et al. 2017) and toxicity (Saveski, Roy, and Roy 2021). We therefore include features that capture different dimensions of a participant's activity. We use average thread values for matching and the raw user-level values in the regression analysis.

[2] All features described in this section were used for Mahalanobis distance matching, unless stated in the description.

	Conflict		*Anger*		*Politeness*		*Sentiment*		*Engagement (log)*	
	Coef	z	Coef	z	Coef	z	Coef	z	Coef	z
Has Conflict	0.044***	(0.000)	-0.000	(0.947)	-0.051***	(0.000)	-0.038***	(0.000)	-0.077***	(0.000)
Prev Conflict	0.006*	(0.021)								
Prev Anger			0.004	(0.099)						
Prev Politeness					0.006*	(0.026)				
Prev Sentiment							0.001	(0.651)		
Prev Engagement (log)									0.590***	(0.000)
User Age	0.005	(0.116)	0.005	(0.090)	0.012***	(0.000)	0.004	(0.192)	0.039***	(0.000)
User Female	-0.005	(0.366)	0.005	(0.359)	0.031***	(0.000)	-0.002	(0.758)	0.009	(0.061)
Friend Count	-0.000	(0.558)	-0.000	(0.340)	-0.000**	(0.008)	0.000	(0.608)	-0.000***	(0.000)
Thread Age (log & std)	-0.002	(0.567)	-0.007*	(0.038)	0.001	(0.812)	-0.000	(0.972)	-0.028***	(0.000)
# Participants in Thread (log)	0.004	(0.236)	0.001	(0.764)	-0.002	(0.438)	0.001	(0.758)	-0.008**	(0.002)
Constant	-0.010	(0.093)	0.015*	(0.013)	0.021***	(0.001)	0.020**	(0.001)	0.031***	(0.000)
Observations	150797		150797		150797		150797		150797	
N_posts	24,275		24,275		24,275		24,275		24,275	
Model Rsq	0.000		0.000		0.001		0.000		0.372	

Table 2: Effects of Conflict on Language and Engagement. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

- *Average number of posts.* The average number of posts participants made during the pre-treatment period.
- *Average number of comments.* The average number of comments participants made during the pre-treatment period.
- *Average number of reactions.* The average number of reactions to posts and comments participants made during the pre-treatment period, where reactions include like, love, sorry, support, wow, anger, and haha.
- *Average number of days of activity in 28 days.* The average number of days participants are active in the group during the 28-day pre-treatment period.

Set 4: Participants' linguistic behavior Prior work finds correlations between language features and whether a conversation or comment contains conflict (Levy et al. 2022; Proksch et al. 2019; Zhang et al. 2018) . We include a comprehensive list of participants' linguistic features as controls. We use average thread values for matching and the raw user-level values in the regression analysis.

- *Politeness.* We calculate the politeness of a thread as the average politeness score based on the BERT-based politeness classifier from Hayati, Kang, and Ungar (2021) run on the comments participants posted in the group during the pre-treatment period. This politeness model was trained on the StanfordPoliteness data (Danescu-Niculescu-Mizil et al. 2013b) and achieved an F1-score of 69.4% on their test data.
- *Sentiment.* Similarly, we calculate the sentiment of a thread as the average positive sentiment probability scores based on Hayati et al's BERT-based sentiment classifier (Hayati, Kang, and Ungar 2021) run on the comments participants posted in the group during the pre-treatment period. This sentiment classifier was trained on the Sentiment Treebank dataset (Socher et al. 2013) with an F1-score of 96.5% on their test data.
- *Anger.* We use a BERT-based anger classifier (Hayati, Kang, and Ungar 2021) trained on a benchmarking emotion data set from SemEval 2018: Affect in Tweets (Mohammad et al. 2018) to collect anger probability scores for our data. The anger classifier is a binary classifier with an F1-score of 82.0% that predicts whether a sentence contains anger. Note we did not use anger for matching, but did confirm that the absolute standard mean differences in anger scores in the final matched sample is small, with an ASMD value of 0.009.
- *Conflict.* For the reported conflict analysis only, we match on the conflict detection model as previously described in the Data and Conflict Detection Model Section.

The practice of applying off-the-shelf models trained in one domain to a new domain is common. A well-known example is the VADER sentiment analyzer (Hutto and Gilbert 2014). Even though it is a rule-based model and only validated on tweet text, it has been widely used (>5000 citations) across domains, such as Facebook, customer service calls, press releases, and marketing communication (e.g. Berger et al. 2020; Hussain et al. 2021). In addition, we should clarify several points about the off-the-shelf models used in our research. First, they were all trained on social media or online community data, so the impact of domain shift on model performance should be minimal. Second, because BERT-based models are designed to capture the associations between language patterns and context and output labels, their performance should not change substantially across domains.

Regression Outcomes Analysis

The substantive goal of the analysis is to examine changes in language behavior and behavioral engagement after participants were exposed to conflict and to conflict reports during the treatment period. For this outcome analysis, we use data from the post-treatment, the 4 weeks after the treatment period. To measure conflict, politeness, sentiment, and anger in participants' posts and comments, we use the pre-trained machine-learning models described previously. If a participant made no posts or comments in the group during the

	Conflict		*Anger*		*Politeness*		*Sentiment*		*Engagement (log)*	
	Coef	z	Coef	z	Coef	z	Coef	z	Coef	z
Reported	0.244***	(0.000)	0.169***	(0.000)	-0.045*	(0.035)	-0.132***	(0.000)	-0.043*	(0.014)
Rep. Commenter	-0.036	(0.153)	-0.039	(0.122)	0.008	(0.767)	0.002	(0.949)	0.075***	(0.000)
Rep. X Rep. Commenter	0.051	(0.152)	0.064	(0.074)	-0.056	(0.116)	-0.025	(0.478)	-0.235***	(0.000)
Prev Conflict	-0.004	(0.620)								
Prev Anger			-0.009	(0.274)						
Prev Politeness					-0.005	(0.584)				
Prev Sentiment							0.008	(0.350)		
Prev Engagement (log)									0.615***	(0.000)
User Age	-0.004	(0.611)	0.006	(0.466)	0.008	(0.333)	-0.002	(0.811)	0.024***	(0.001)
User Female	-0.032	(0.063)	-0.010	(0.556)	0.018	(0.305)	0.014	(0.408)	0.023	(0.105)
Friend count	-0.000	(0.294)	-0.000	(0.542)	-0.000	(0.638)	-0.000	(0.378)	-0.000	(0.861)
Thread Age (log & std)	-0.004	(0.653)	0.001	(0.889)	0.003	(0.755)	-0.013	(0.134)	-0.027**	(0.001)
# Participants in Thread (log)	0.023**	(0.009)	0.011	(0.192)	-0.012	(0.166)	-0.015	(0.089)	-0.013	(0.118)
Constant	-0.092***	(0.000)	-0.072***	(0.000)	0.023	(0.236)	0.067***	(0.000)	0.024	(0.129)
Observations	13679		13679		13679		13679		13679	
N_posts	3,914		3,914		3,914		3,914		3,914	
Model Rsq	0.018		0.009		0.002		0.006		0.404	

Table 3: Effects of Conflict Reports & User Type on Language & Engagement. Rep. stands for Reported. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

observation period, we substitute the median value in the group for the missing values. However, a robustness check not reported for reasons of space shows qualitatively similar results if we use multiple imputation to substitute for missing values (Rubin 2004; Schafer 1999). To measure participants' behavioral engagement in the group, we sum the numbers of posts, comments, and reactions they made in the group during the observation period. Posts are images and/or text intended to start a conversation, comments are replies to one's own or others posts and comments, and reactions are single-click icons that participants can apply to a post or comment. We log these engagement measures before taking their average.

We use linear regression to estimate the effects of conflicts, reports, user type, and the control variables on the language and behavioral engagement outcomes. Because of the long-tailed nature of the behavioral engagement measure, we also computed a negative binomial regression, appropriate for count data, as a robustness check. These results are reported in Appendix Table 4. The units of analysis are the individual users who participated in the target conversations during the treatment period. Because users who participated in the same thread are not independent of each other, we use random-effects, hierarchical regression models, with users nested within the thread, to account for their non-independence. The input variables for these models include those described in the Matching Section in addition to binary labels designating whether a conversation contained conflict, whether conflict was reported, and whether a participant was the target of the conflict report.

To estimate the impact of being involved in a conflict thread versus a non-conflict one, we include in the regression model the '*Has Conflict*' binary feature set to 1 if the user is a participant in a conversation containing conflict and 0 otherwise. Similarly, to estimate the impact of conflict reports, we include a '*Reported*' binary feature that is set to 1 if the participant is involved in a conflict thread that was reported to an administrator and set to 0 if the conflict was not reported. Deterrence theory differentiates specific deterrence, which applies to people who were the target of an administrative action, from general deterrence, which applies to others learning about the administrative action. Similarly, learning theories differentiate the effects of individual feedback from feedback to a group. Therefore, when testing the impact of conflict reports we differentiate the person who was the target of a report ($Rep.Commenter = 1$) from other participants in the conversation who were not targets ($Rep.Commenter = 0$). To determine whether reports have different effects on the reported commenter versus others in the conversation, we include the statistical interaction between reports and reported commenter (i.e., $Rep.X Rep.Commenter$) in the regression analyses.

Results and Discussion

Table 2 shows the results for the regression analyses predicting participants' language and behavioral engagement in the group in post-treatment observation period depending on whether or not they were exposed to a conflict thread during the treatment period. Consistent with theories of behavioral mimicry and with H1, participants who were in a conversation with conflict during the treatment period expressed .04 standard deviations more conflict in their posts and comments during the post-treatment observation period than those who were not exposed to conflict ($\hat{\beta} = .044, p < .001$). Consistent with H2, they were also less polite in their language ($\hat{\beta} = -.051, p < .001$) and showed less positive sentiment ($\hat{\beta} = -.038, p < .001$). They did not differ in the amount of anger in their language ($\hat{\beta} = .000, p > .05$). Consistent with H3, participants exposed to a conflict became less engaged in the group, creating fewer posts, comments, and reactions during the post-treatment observation

period, compared to those were not exposed to conflict ($\hat{\beta} = -.077, p < .001$). While the results in Table 2 indicating lower engagement were based on a random-effects linear regression analysis of a log transformed measure of engagement, they are consistent with results using a negative binomial regression analysis. This shows that user's rate of posts, comments, and reactions decrease by .87 when they are exposed to a conflict comment.

Table 3 shows the results for regression analyses predicting participants' language and behavioral engagement depending on whether the conflict thread that they were involved in was reported to an administrator or not and whether the participant was the target of the report. *Inconsistent* with theories of both specific and general deterrence and H4a and H5, participants involved in a conflict conversation reported to administrators during the treatment period expressed *more* conflict in their subsequent posts and comments during the post-treatment observation period than those who were exposed to conflict that was not reported ($\hat{\beta} = .244, p < .001$). Their language was also less polite ($\hat{\beta} = -.045, p < .001$), showed less positive sentiment ($\hat{\beta} = -.132, p < .001$) and more anger ($\hat{\beta} = .169, p < .001$). These effects did not differ for the reported participant or others in the conversation, with all (*P*-values for the Reported X Reported Commenter interactions for the language outcomes being less than 0.05). It is not clear why all the language indicators of positive tone in the group deteriorated after conflict was reported to an administrator. If the reason is that the person reported was upset, one would expect that the deterioration in language tone would be greater for the instigator than others in the conversation, but statistical analyses showed no reliable differences in the language used by the target of the report and others. Indeed many people in the group, including the person instigating the conflict, might not be aware of the report, since in many cases administrators may hide the conflict comment from the group or ban the instigator without telling anyone. Another possibility is that the reports made more salient the conflictful behavior and thereby accentuated the negative effects of conflict on the outcomes. Finally, a more methodological explanation is that conflict conversations that were reported generally had more negative language than conversations that were not reported, but that our matching procedure did not sufficiently control for the pre-existing differences in language.

Although other people in the conversation engaged in the group slightly less when the conflict was reported to an administrator ($\hat{\beta} = -.043, p < .05$), consistent with theories of specific deterrence and H4b the decline in engagement was substantially larger for the person who had been reported (for the Reported X Reported Commenter interaction $\hat{\beta} = -.235, p < .05$). This large drop in behavioral engagement among those whose conflict was reported could have occurred because feedback from administrators, including coaching or chastisements, could have caused the instigator of conflict to improve his or her behavior. However, another possibility is that administrators may have directly intervened to reduce the instigators' ability to disrupt conversations by banning the instigators for a period or reducing the visibility of their subsequent posts and comments. Unfortunately, our data does not include information about the administrator's response to a report.

Conclusion and Future Work

Discussion. The goal of this research was to explore the downstream consequences of conflict in online groups and the reporting of the conflict on the language people subsequently used in the group and their subsequent behavioral engagement in the group. To do so, this paper contrasts online conversations in Facebook groups in which conflict occurred to similar conversation where little conflict occurred. It also contrasts conflictful conversations where the conflict was reported to a group administrator to similar ones where the conflict was not reported. If we assume that these contrasts reflect causation, then conflict in group conversations has bad consequences. It degrades the tone of the conversation, leading to language containing more conflict, angry and negative language with less politeness, and leading people in the conversation to reduce their level of participation in the group. These results are in general consistent with theories of imitation and social mimicry. When people see conflict, they follow-up by using more negative language, and the resulting degradation of tone in the group may lead people to reduce their participation in the group. Although the effects of a single conflictful conversation are very small, with the mean absolute effect size being only 0.04 standard deviations, they are substantively meaningful because of the large number of conflicts people might be exposed to when participating in some groups over time and because the effects of conflict might be cumulative.

Reporting the conflict seems to have mixed effects, encouraging more negative language in the group (mean absolute effect size of 0.33 standard deviations) but at the same time protecting the group, by leading the instigators of the conflict to participate less in the group, either voluntarily or because the administrators curtailed their participation by banning them or reducing the visibility of their posts and comments.

Limitations. However, the causation assumption, which is the basis of this reasoning, may not be correct. The major limitation in this research is its observational nature. Unlike random-assignment experiments, where people are randomly exposed to conflict or not or where instigators of conflict are randomly reported to administrators or not, the data from this study are observational. We used Mahalanobis distance matching (MDM) to try to approximate a random-assignment experiment. An examination of the absolute mean differences across many potential confounds indicates the matching substantially reduced pre-existing differences among participants exposed to conflict or not and between those in threads in which conflict was reported or not. However, the pre-existing differences were not completely eliminated. Moreover, although we matched on many potential confounds, including by matching within groups to control for group characteristics and by matching on characteristics of participants such as their demographics, characteristics of the conversations such as the number of people involved, and the outcome variables measured the month

prior to the treatment period, we undoubtedly failed to match on all potential confounds. In addition, despite the popularity of matching procedures to draw causal inferences from observational data, some methodologists argue that under some circumstances these pruning techniques, which conduct analyses on a matched subset of data, can paradoxically increase imbalance between treatment and control groups in non-experimental settings (King and Nielsen 2019). Additionally, due to the small number of conflicts per group, MDM matching becomes infeasible if we want to match within a group. Thus, it is possible that some of the observed results may be the result of partial failures in the matching procedures we used and reflect pre-existing differences among users rather than the effects of conflicts and reports. Finally, we acknowledge limitations from relying on user-reported conflict as ground-truth labels.

Future Work and Practical Implications. Future work could consider heterogeneity in conflict and reporting effects by group interest. For example, the consequences of conflict and reports of conflict may differ for sports groups versus political ones. Next, extensions could characterize effective conflict reports that reduce incidents of conflict without reducing participants' engagement. Finally, while this work is focused on short-term consequences of conflict, one could evaluate longer-term effects.

A/B experiments often used for product testing within social media companies are needed to more definitively determine the causal impact of conflicts in online groups and the effects of reporting of conflict. Although it would be unethical to artificially generate conflict in a group, it would be possible and ethical to use tools like the conflict-detection algorithm developed for this research to reduce the visibility of conflictful comments and examine their consequences on those who see them compared to those who do not.

Similarly, these algorithms could be used to automatically alert administrators to the presence of conflict in their groups, although how they respond could be based on their discretion. To more precisely determine the impact of different conflict mitigation strategies while simultaneously retaining administrator agency and keeping human judgment in the loop, these alerting tools could be paired with improved tools to help administrators take appropriate courses of action, such as hiding the conflictful comment from other group members or communicating with the instigator of the conflict.

Given recent large improvements in the development of large language models in natural language applications such as chatbots for improved mental health (Crasto et al. 2021) or argument mining (Habernal et al. 2023), even more capable and complex conflict detection can be trained than the one used in this work. Additionally, these advanced tools could be leveraged to provide administrators a summary of the conflict, to lighten their burden, or recommendations on ways to mitigate the conflict or prevent future ones. Prior research has used language models to reduce toxicity and improve conversation quality (Argyle et al. 2023), and it is possible to build a system to generate recommendations to reduce conflict. Rather than focusing only on administrators, a conceptually similar system be deployed for any group member as an assistant to provide suggestions on how to de-escalate a conflict through an apology or a clarification of a misunderstanding.

References

Abedin, E.; Jafarzadeh, H.; and Akhlaghpour, S. 2018. Opinion mining on Twitter: A sentiment analysis of the Iran deal.

Akers, R. L.; and Cochran, J. K. 1985. Adolescent marijuana use: A test of three theories of deviant behavior. *Deviant Behavior*, 6(4): 323–346.

Akers, R. L.; and La Greca, A. J. 1991. Alcohol use among the elderly: Social learning, community context, and life events. *Society, Culture, and Drinking Patterns Reexamined*, 1: 242.

Allred, K. G.; Mallozzi, J. S.; Matsui, F.; and Raia, C. P. 1997. The influence of anger and compassion on negotiation performance. *Organizational Behavior and Human Decision Processes*, 70(3): 175–187.

Andersson, L. M.; and Pearson, C. M. 1999. Tit for tat? The spiraling effect of incivility in the workplace. *Academy of Management Review*, 24(3): 452–471.

Argyle, L. P.; Bail, C. A.; Busby, E. C.; Gubler, J. R.; Howe, T.; Rytting, C.; Sorensen, T.; and Wingate, D. 2023. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41): e2311627120.

Arpino, B.; and Cannas, M. 2016. Propensity score matching with clustered data. An application to the estimation of the impact of caesarean section on the Apgar score. *Statistics in Medicine*, 35(12): 2074–2091.

Asch, S. E. 1956. Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, 70(9): 1.

Austin, P. C. 2009. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, 28(25): 3083–3107.

Bakshy, E.; Rosenn, I.; Marlow, C.; and Adamic, L. 2012. The role of social networks in information diffusion. In *Web Conference*.

Bandura, A.; and Walters, R. H. 1977. *Social Learning Theory*, volume 1. Prentice Hall.

Berger, J.; Humphreys, A.; Ludwig, S.; Moe, W. W.; Netzer, O.; and Schweidel, D. A. 2020. Uniting the tribes: Using text for marketing insight. *Journal of Marketing*, 84(1): 1–25.

Bodtker, A. M.; and Jameson, J. K. 2001. Emotion in conflict formation and its transformation: Application to organizational conflict management. *International Journal of Conflict Management*.

Bruckman, A. 2002. Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet. *Ethics and Information Technology*, 4(3): 217–231.

Chambers, N.; Bowen, V.; Genco, E.; Tian, X.; Young, E.; Harihara, G.; and Yang, E. 2015. Identifying political sentiment between nation states with social media. In *EMNLP*, 65–75.

Chartrand, T. L.; and Bargh, J. A. 1999. The chameleon effect: the perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6): 893.

Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2015. Antisocial behavior in online discussion communities. In *Proceedings of the international aaai conference on web and social media*, volume 9, 61–70. ISBN 2334-0770.

Chilton, P. 1990. Politeness, politics and diplomacy. *Discourse & Society*, 1(2): 201–224.

Cochran, W. G.; and Rubin, D. B. 1973. Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, 417–446.

Coletto, M.; Garimella, K.; Gionis, A.; and Lucchese, C. 2017. Automatic controversy detection in social media: A content-independent motif-based approach. *Online Social Networks and Media*, 3: 22–31.

Crasto, R.; Dias, L.; Miranda, D.; and Kayande, D. 2021. CareBot: A Mental Health ChatBot. In *2021 2nd international conference for emerging technology (INCET)*, 1–5. IEEE.

Danescu-Niculescu-Mizil, C.; Sudhof, M.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013a. A computational approach to politeness with application to social factors. *arXiv:1306.6078*.

Danescu-Niculescu-Mizil, C.; Sudhof, M.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013b. A computational approach to politeness with application to social factors. In *ACL*, 250–259. Sofia, Bulgaria: ACL.

De Dreu, C. K. 2010. Social conflict: The emergence and consequences of struggle and negotiation. In Fiske, S. T.; Gilbert, D. T.; and Lindzey, G., eds., *Handbook of Social Psychology*, 983–1023. NY: John Wiley Sons.

De Dreu, C. K. W.; and Weingart, L. R. 2003. Task versus relationship conflict, team performance, and team member satisfaction: A meta-analysis. *Journal of Applied Psychology*, 88(4): 741–749.

de la Vega, L. G. M.; and Ng, V. 2018. Modeling trolling in social media conversations. In *Conference on Language Resources and Evaluation (LREC 2018)*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Association for Computational Linguistics*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dölling, D.; Entorf, H.; Hermann, D.; and Rupp, T. 2009. Is deterrence effective? Results of a meta-analysis of punishment. *European Journal on Criminal Policy and Research*, 15(1): 201–224.

Gibbs, J. P. 1985. Deterrence theory and research. *Nebraska Symposium on Motivation*, 33: 87–130.

Habernal, I.; Faber, D.; Recchia, N.; Bretthauer, S.; Gurevych, I.; Spiecker genannt Döhmann, I.; and Burchard, C. 2023. Mining legal arguments in court decisions. *Artificial Intelligence and Law*, 1–38.

Hayati, S. A.; Kang, D.; and Ungar, L. 2021. Does BERT Learn as Humans Perceive? Understanding Linguistic Styles through Lexica. In *Conference on Empirical Methods in Natural Language Processing*.

Higgins, G. E.; Wilson, A. L.; and Fell, B. D. 2005. An application of deterrence theory to software piracy. *Journal of Criminal Justice and Popular Culture*, 12(3): 166–184.

Hussain, A.; Tahir, A.; Hussain, Z.; Sheikh, Z.; Gogate, M.; Dashtipour, K.; Ali, A.; and Sheikh, A. 2021. Artificial intelligence–enabled analysis of public attitudes on facebook and twitter toward covid-19 vaccines in the united kingdom and the united states: Observational study. *Journal of medical Internet research*, 23(4): e26627.

Hutto, C.; and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, 216–225.

Kasper, G.; Blum-Kulka, S.; Pragmatics, I.; Bialystok, E.; Realization, I. S. A.; Bergman, M. L.; Hints, I. R.; and Zuengler, J. 1993. Oxford University Press, Inc.

King, G.; and Nielsen, R. 2019. Why propensity scores should not be used for matching. *Political Analysis*, 27(4): 435–454.

King, G.; Nielsen, R.; Coberley, C.; Pope, J. E.; and Wells, A. 2011. Comparative effectiveness of matching methods for causal inference. *Unpublished Manuscript*.

Kluger, A. N.; and DeNisi, A. 1996. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2): 254.

Kujur, F.; and Singh, S. 2018. Emotions as predictor for consumer engagement in YouTube advertisement. *Journal of Advances in Management Research*.

Lerner, J. S.; and Keltner, D. 2001. Fear, anger, and risk. *Journal of Personality and Social Psychology*, 81(1): 146.

Levy, S.; Kraut, R. E.; Yu, J. A.; Altenburger, K. M.; and Wang, Y.-C. 2022. Understanding Conflicts in Online Conversations. In *Web Conference*. NY: ACM.

Lucić, D.; Katalinić, J.; and Dokman, T. 2020. Sentiment analysis of the syrian conflict on twitter. *Media Studies*, 11(22): 46–61.

Ma, X.; Cheng, J.; Iyer, S.; and Naaman, M. 2019. When do people trust their social groups? In *CHI*, 1–12.

Mahalanobis, P. C. 1936. On the generalized distance in statistics. National Institute of Science of India.

Maimon, D.; Howell, C. J.; and Burruss, G. W. 2021. Restrictive deterrence and the scope of hackers' reoffending: Findings from two randomized field trials. *Computers in Human Behavior*, 125: 106943.

Mohammad, S.; Bravo-Marquez, F.; Salameh, M.; and Kiritchenko, S. 2018. SemEval-2018 Task 1: Affect in Tweets. In *Workshop on Semantic Evaluation*, 1–17. New Orleans: ACL.

Nelson-Field, K.; Riebe, E.; and Newstead, K. 2013. The emotions that drive viral video. *AMJ*, 21(4): 205–211.

Obeng, S. G. 1997. Language and politics: Indirectness in political discourse. *Discourse & Society*, 8(1): 49–83.

Perez, S. 2021. Facebook rolls out new tools for group admins, including Automated Moderation AIDS.

Phillips, D. P. 1974. The influence of suggestion on suicide: Substantive and theoretical implications of the Werther effect. *American Sociological Review*, 340–354.

Proksch, S.-O.; Lowe, W.; Wäckerle, J.; and Soroka, S. 2019. Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. *Legislative Studies Quarterly*, 44(1): 97–131.

Romero, D. M.; Meeder, B.; and Kleinberg, J. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. In *Web Conference*.

Rosenbaum, P. R.; and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55.

Rubin, D. B. 1979. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *JASA*, 74(366a): 318–328.

Rubin, D. B. 2004. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons. ISBN 0471655740.

Saveski, M.; Roy, B.; and Roy, D. 2021. The structure of toxic conversations on Twitter. In *Web Conference*.

Schafer, J. L. 1999. Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1): 3–15.

Seering, J.; Kraut, R. E.; and Dabbish, L. 2017. *Shaping pro and anti-social behavior on Twitch through moderation and example-setting*. NY: ACM.

Sekhon, J. S. 2009. Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science*, 12(1): 487–508.

Sharma, S.; Elfenbein, H. A.; Sinha, R.; and Bottom, W. P. 2020. The effects of emotional expressions in negotiation: a meta-analysis and future directions for research. *Human Performance*, 33(4): 331–353.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A.; and Potts, C. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *EMNLP*.

Sood, S. O.; Churchill, E. F.; and Antin, J. 2012. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology*, 63(2): 270–285.

Stafford, M. C.; and Warr, M. 1993. A reconceptualization of general and specific deterrence. *Journal of Research in Crime and Delinquency*, 30(2): 123–135.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 25(1): 1.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web (WWW 201)*, 1391–1399. Association for Computing Machinery.

Xu, B.; Xu, Z.; and Li, D. 2016. Internet aggression in online communities: a contemporary deterrence perspective. *Information Systems Journal*, 26(6): 641–667.

Yeomans, M.; Minson, J.; Collins, H.; Chen, F.; and Gino, F. 2020. Conversational receptiveness: Improving engagement with opposing views. *Organizational Behavior and Human Decision Processes*, 160: 131–148.

Yildirim, M. M.; Nagler, J.; Bonneau, R.; and Tucker, J. A. 2023. Short of suspension: How suspension warnings can reduce hate speech on twitter. *Perspectives on Politics*, 21(2): 651–663.

Zhang, J.; Chang, J. P.; Danescu-Niculescu-Mizil, C.; Dixon, L.; Hua, Y.; Thain, N.; and Taraborelli, D. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of ACL*.

Zhu, H.; Kraut, R.; and Kittur, A. 2012. Organizing without formal organization: group identification, goal setting and social modeling in directing online production. In *CSCW*, 935–944. NY: ACM.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? yes
   (e) Did you describe the limitations of your work? yes
   (f) Did you discuss any potential negative societal impacts of your work? yes
   (g) Did you discuss any potential misuse of your work? yes
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? yes
   (b) Have you provided justifications for all theoretical results? yes

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? yes

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? yes

(e) Did you address potential biases or limitations in your theoretical framework? yes

(f) Have you related your theoretical results to the existing literature in social science? yes

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? yes

3. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? yes

(b) Did you mention the license of the assets? n/a

(c) Did you include any new assets in the supplemental material or as a URL? no

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? yes

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? yes

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? n/a

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? n/a

Appendix

	Conflict		*Reports*	
	Coef	z	Coef	z
Has Conflict	-0.135***	(-21.77)		
Conflict Rep.			-0.059*	(-2.57)
Rep. Commenter			0.133***	(5.33)
Conflict Rep.X Rep. Commenter			-0.482***	(-12.89)
Prev Engagement	0.000***	(-139.47)	0.001***	(53.61)
User Age	0.094***	(-34.02)	0.067***	(7.28)
User Female	0.040***	(-6.92)	0.049*	(2.57)
Friend Count	-0.000***	(-8.64)	0	(0.34)
Thread Age (std)	-0.033***	(-10.12)	-0.052***	(-5.22)
# Participants	-0.011***	(-3.60)	0.001	(0.07)
Constant	-0.649***	(-100.83)	-0.973***	(-42.86)
Observations	150797		13679	
ln_r	0.097***	0	0.749***	0
ln_s	4.233***	0	5.527***	0

Table 4: Negative Binomial Regression: Effects of Conflict, Conflict Reports, & User Type on Engagement. Prev Engagement, Thread Age, and # Participants are log transformed. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$