Understanding Conflicts in Online Conversations

Sharon Levy*
University of California, Santa Barbara
Santa Barbara, CA, USA
sharonlevy@cs.ucsb.edu

Robert E. Kraut
Carnegie Mellon University
Pittsburgh PA, USA
robert.kraut@cmu.edu

Jane A. Yu
Meta
Menlo Park, CA, USA
janeyu@fb.com

Kristen M. Altenburger
Meta
Menlo Park, CA, USA
kaltenburger@fb.com

Yi-Chia Wang
Meta
Menlo Park, CA, USA
yichiaw@fb.com

ABSTRACT

With the rise of social media, users from across the world are able to connect and converse with each other online. While these connections have facilitated a growth in knowledge, online discussions can also end in acrimonious conflict. Previous computational studies have focused on creating online conflict detection models from inferred labels, primarily examine disagreement but not acrimony, and do not examine the conflict's emergence. Social science studies have investigated offline conflict, which can differ from its online form, and rarely examines its emergence. The current research aims to understand how online conflicts arise in online personal conversations. Our ground truth is a Facebook tool that allows group members to report conflict to administrators. We contrast discussions ending with a conflict report with paired non-conflict discussions from the same post. We study both user characteristics (e.g., historical user-to-user interactions) and conversation dynamics (e.g., changes in emotional intensity over the course of the conversation). We use logistic regression to identify the features that predict conflict. User characteristics such as the commenter's gender and previous involvement in negative online activity are strong indicators of conflict. Conversational dynamics, such as an increase in person-oriented discussion, are also important signals of conflict. These results help us understand how conflicts emerge and suggest better detection models and ways to alert group administrators and members early on to mediate the conversation.

CCS CONCEPTS

• **Networks** → **Social media networks**; • **Information systems** → **Social networks**; • **Computing methodologies** → **Natural language processing**; • **Applied computing** → *Sociology*.

KEYWORDS

online conflicts, conversation analysis, natural language processing

*This research was done while interning at Meta.

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9096-5/22/04.
https://doi.org/10.1145/3485447.3512131

ACM Reference Format:
Sharon Levy, Robert E. Kraut, Jane A. Yu, Kristen M. Altenburger, and Yi-Chia Wang. 2022. Understanding Conflicts in Online Conversations. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3485447.3512131

1 INTRODUCTION

Interpersonal conflicts disrupt effective communication both online and in offline daily life. Social science literature routinely distinguishes between substantive task-related conflict, in which people disagree about procedures and goals, and affective interpersonal conflict, in which the conflict revolves around the relations of the participants and their frustration and dislike of each other [13, 19, 24]. Substantive conflict can positively impact people who engage in constructive debates, by encouraging them to discuss alternative perspectives, reconcile differences, and develop empathy for different viewpoints. However, substantive disagreements that morph into interpersonal or relationship conflict generally have negative consequences for both individuals and groups. Relationship conflict is strongly related to reductions in trust, cohesion, satisfaction, commitment, voluntary good citizenship behavior, and positive affect in work groups and teams [13]. It can also result in biased information processing [14] and poorer quality work [11] . Interpersonal conflicts may arise in both online and offline settings and can evolve from differences in goals, opinions, or group identification. In addition to these potential sources of conflict, the rapid turnover in membership in online groups, lean communication media (i.e., media without additional information from non-verbal cues) that may lead members to forget they are communicating with other humans, and the existence of trolls, suggest that interpersonal conflict may be more common online than offline and may differ in its causes [12]. As a result, the study of online interpersonal conflict is important in its own right as well as providing a lens to study conflict escalation processes more generally. Although some research on online conflict has analyzed more neutral disagreements [21], or focused on specific types of harmful conflict such as trolls deliberately posting provocative or offensive comments to provoke others [12], we study the rare non-constructive, interpersonal conflicts (< 0.0001% in our data set), which escalate and lead users to report the conflict to their group administrators.

Prior work on online conflicts has primarily focused on creating conflict detection models. These models are trained to detect

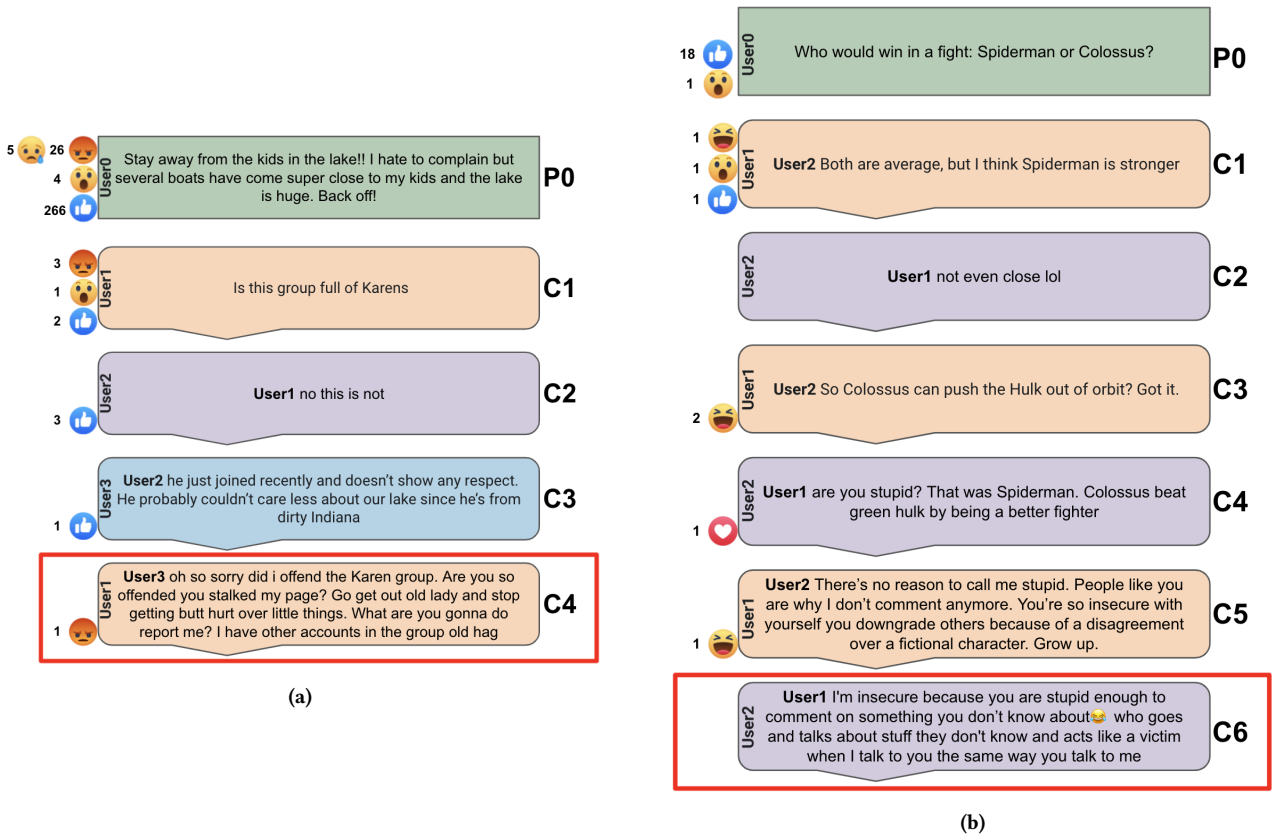


Figure 1: Examples of conflict conversations, where each text sequence is edited and paraphrased. The original, initiating post P0 (also edited and paraphrased) is shown at the top and reply-comments (C1-C6), which constitute the conversation thread, are shown below. The comments in the red boxes were reported to the group administrators as conflicts.

conflicts at the post-level or topic-level [1, 16, 18], with the implication that all of the conversations associated with these posts and topics are either conflictful or not. Additionally, these studies infer whether a comment reflects conflict from metadata such as upvote/downvote ratios and topic of discussion [9, 21, 27]. As a result, this research does not differentiate substantive disagreements from more toxic interpersonal conflict. Most studies of conflict in the social sciences have evaluated conflict in offline settings, examining relatively static causes (e.g., differences in interests among participants), downstream effects of the conflict (e.g., work groups cohesion or performance), and types of conflict resolution (e.g., mediation strategies) [23, 33, 46]. However, little research investigating conflict in online or offline settings has examined how interpersonal conflict emerges and the factors that lead to it [26, 44].

The focus of our paper is on the dynamics of online interpersonal conflicts. Online conflicts can differ from other conflicts multiple ways due to the nature of social media technology. First, there is a wide geographical distribution across users, which can spawn cultural differences leading to conflict [43]. Additionally, online users often have conversations with strangers [20] or easily join existing conversations, allowing for an increase in the number of participants and the likelihood that people are interacting with strangers. Finally, in online groups, turnover is high, the communication media are lean, and participants perceive themselves to be relatively anonymous compared to offline interactions. These characteristics may result in volatile conversations that are more dynamic in the number of participants and opinions. While this can allow users to engage with each other and discuss differing viewpoints respectfully, it can also lead to escalated disagreements among users. As the prior research in both computer science and social science has not studied the emergence of online conflict, there is a gap in current knowledge in this area.

We aim to fill this gap and contribute to research on the emergence of online interpersonal conflict. To do this, we take advantage of Facebook's conflict comment reporting tool[1], in which members of a group can report a comment with objectionable conflict to a group administrator. We contrast conflict conversations and non-conflict ones on the same topic in the same group. Figure 1 shows two paraphrased examples of conversations with a reported conflict comment. We utilize the data to explain the factors that lead to a conflict in Facebook group conversations and distinguish pre-existing *user characteristics*, which participants bring into the

[1]https://www.facebook.com/community/whats-new/new-tools-features-nurture-community/

Table 1: Hypotheses and their respective categories.

User Characteristics	H1	Conflict commenters have pre-existing differences from other commenters
	H2	Conflict commenters are more likely to be involved in other negative activity online
	H3	Conflict commenters are not well-connected to the group where the conflict occurs
Conversation Dynamics	H4	Conflict threads have an increase in intensity and negative emotions
	H5a	Conflict conversations are identity/group-based rather than interpersonal
	H5b	Conflict conversations become more interpersonal over time

conversation, and *conversation dynamics*, which emerge over the course of the conversation. These features are inputs to a multilevel logistic regression model to predict whether the conversation ends in reported conflict. Our results promote theoretical understanding of how conflicts arise in a discussion, extending other work on the dynamics of conflict evolution [44]. In addition, they can be used practically to build better conflict detection models and alert systems to notify group administrators and moderators and even participants in the conversation of an emerging conflict for early intervention.

Our contributions include: 1) using ground truth labels from the people involved to distinguish conflict and non-conflict conversations, 2) empirically identifying variables that may lead to conflict within online conversations as the conflict is emerging, and 3) distinguishing static characteristics of users and dynamic characteristics of the conversation, including how it changes over time, to explain the emergence of interpersonal conflict.

2 RELATED WORK AND HYPOTHESES

In this section, we review related literature in the area of conflict and develop hypotheses for our analysis on the emergence of online interpersonal conflict.

Varying demographic characteristics can lead to differing engagement with conflict, as shown in prior research. For example, Eagly and Steffen [17] showed that women are somewhat less aggressive than men, and Holt and DeVore [23] found that females are more likely to compromise than males during disagreements, which can diffuse the situation rather than escalate it further. Triandis [43] demonstrated the existence of cultural and racial differences in conflict. Cheng et al. [6] showed that antisocial behavior in online groups is a relatively stable individual difference. Together, this research suggests that a number of pre-existing user characteristics may lead to online, interpersonal conflict:

- **H1**: Conflict commenters have pre-existing differences from other commenters.

Related work in the area of toxic conversations has evaluated the connections between the users involved in the discussions. Saveski et al. [36] found that the largest proportion of toxic tweets in their study was posted by moderately toxic users (users who have posted several toxic tweets), demonstrating a high probability of repeat offenders. This leads us to believe conflict commenters may be involved in other such activities online, formulating our next hypothesis:

- **H2**: Conflict commenters are more likely to be involved in other negative activity online.

Additionally, Saveski et al. [36] showed that toxic replies in conversations occur from users with weaker social connections and fewer friends in common with the poster. Coletto et al. [9] similarly studied connections of users in the context of controversial threads. To do so, they analyzed local network patterns of user-follower and user-reply graphs. Their findings showed controversial interactions are less likely to occur between users who follow each other on social media. With this in mind, we formulate our third hypothesis:

- **H3**: Conflict commenters are not well-connected to the group where the conflict occurs.

Previous research on controversy detection has utilized both sentiment and emotions to detect the extent to which a topic or discussion is controversial [1, 7, 27, 30, 42]. These papers have analyzed intensity of emotions and sentiments, showing that these text-based features are strong indicators of controversy. Meanwhile, Coletto et al. [9] investigated intensity through a different lens, by assuming controversial topics generate "dense" discussions, so that the inter-reply rate for these conversations are lower (i.e., more rapid replies) than those of a non-controversial topic. Zhang et al. [47] studied the early derailment of conversations within Wikipedia talk page discussions through a variety of text-based features such as politeness strategies and prompt types. On the social science side, Weingart et al. [44] analyzed offline conflicts in terms of conflict spirals, characterized by escalating tension in the conflict conversations and increases in reciprocated negative communications (e.g., threats) and emotional states. While previous work examined overall sentiment/emotional intensity online or examined changes in it in an offline setting, we predict a pattern of increasing negative emotions and intensity in online conversations. This leads us to our next hypothesis:

- **H4**: Conflict threads have an increase in intensity and negative emotions.

de Dreu [10] argued that intergroup conflicts are more common than interpersonal ones. Many of these conflicts are identity-based, in which people believe that the group or subgroup with which they identify (e.g., ethnicity, race, religion, or political party) is superior to an outgroup [22, 38, 43]. While many conflicts have their origins in intergroup differences, these intergroup conflicts can evolve to become interpersonal ones [28], although frequent pleasant interactions can reduce intergroup conflicts [32]. The previous literature suggests two hypotheses:

- **H5a**: Conflict conversations are identity/group-based rather than interpersonal.
- **H5b**: Conflict conversations become more interpersonal over time.

We categorize our hypotheses into two groups: *user characteristics* and *conversation dynamics*. H1, H2, and H3 examine whether online conflicts in groups can be attributed to pre-existing user characteristics, such as their demographics or involvement in other online activity. This enables us to determine which user characteristics reflect a user's propensity to engage in conflict. In contrast, H4 and H5 examine how conflicts emerge during a conversation and whether these conversations differ from non-conflict conversations in relation to emotional intensity and group identity. We summarize our hypotheses and their respective groupings in Table 1.

3 DATA COLLECTION

To collect our data, we utilize Facebook's conflict reporting tool, which allows group members to report comments containing conflict to group administrators for action. Using this tool, people can select "Report comment to group admins" from a drop-down menu next to a comment. This leads to a pop-up menu which allows users to select a reason for reporting the comment: Breaks Group Rule, Fake News, Member Conflict, Spam, Harassment, Hate Speech, Nudity or Sexual Activity, Violence, or Other. A visualization of this tool can be seen in Figure 5 in the appendix. The interface only provides the name of the menu item, "Member Conflict", but does not provide a definition or additional details about the type of comments that should be reported as conflict. Our sample of conflict comments consists of de-identified comments from public groups (any size) and large private groups (>32 members) written in English from 05/29/2021 to 08/15/2021 that were reported by a member as containing conflict.

Because our goal is to understand the differences between conflict and non-conflict conversations, we collected a matched sample of non-conflict conversations by randomly sampling a comment not reported as containing conflict from a separate discussion thread under the same post. We restricted the randomly sampled comment to have a minimum depth of 4 in the conversation thread, which is the minimum depth for reported comments in our analysis. In the remainder of this paper, we refer to conflict comments and non-conflict comments as *target* comments. Since our dataset consists of paired samples from the same post, we removed any target comments written by users who appear as both conflict and non-conflict commenters, to ensure there is no overlap between users in the conflict and non-conflict sets. [2] To study the conversation history, we reconstruct the conversations in which the target comments appeared by retrieving all comments leading up to the target comment, starting at the top-level comment for the post. The final dataset consists of 15,438 paired conflict and non-conflict conversational threads from 10,179 different groups. The data collection process is illustrated in Figure 2.

Previous work on conflicts have used the controversial nature of the topic of discussion as an indicator of conflict [9, 15, 18]. Our data collection process ensures that paired samples of conflict and non-conflict comments discuss the same topic because they are follow-ups to the same initiating post. In addition to controlling for topic, this approach allows us to control for the group's culture and member distribution. Although we are unable to study how topics and group characteristics influence conflict with this approach, our analysis can focus on the user and conversational differences that lead to conflict.

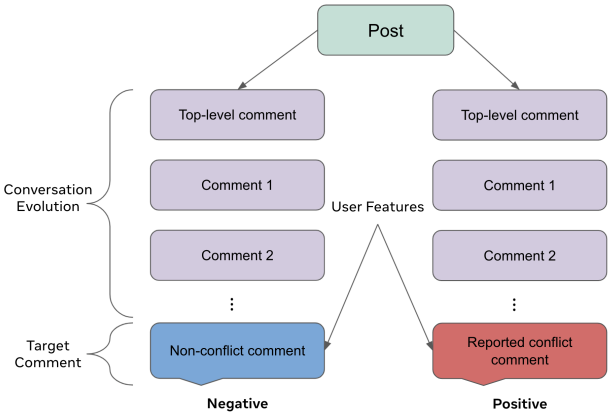


Figure 2: Data collection process for conflict and non-conflict conversations.

4 METHODOLOGY

4.1 Modeling

We use logistic regression to predict whether a target comment is a conflict or non-conflict based on user characteristics and conversation dynamics. Because the paired conflict and non-conflict comments are not independent of each other, we use random intercept, multilevel logistic regression with comments nested under post ID. We build seven models for our analysis, where each model includes an additional set of variables to test subsequent hypotheses (e.g., hypotheses testing the influence of conversational dynamics include all the user characteristics). Results are reported in terms of the odds ratio for each variable, indicating its association with conflict. All continuous variables were standardized with a mean of 0 and standard deviation of 1 before input[3]. Thus, the odds ratio indicates the extent to which changing a dichotomous variable from zero to one (e.g., female to male) or changing a continuous variable by a standard deviation influences the odds of the comment having been reported as conflict. In addition to the odds ratios, we report McFadden's R^2, BIC, Log Likelihood and Log Likelihood Ratio test values for each model. The latter is calculated between pairs of models with an additional set of features (e.g. H2 and H3). We discuss how we operationalize each feature and hypothesis below. The appendix contains descriptive statistics for each of the variables used in the logistic regression.

4.2 User Characteristics

To test whether pre-existing user characteristics predict conflict, we examine the differences between the commenter of the conflict target comment versus the commenter of the non-conflict target comment. To test whether conflict and non-conflict commenters differ on relatively static user characteristics (**H1**), we compare

[2]This mainly occurred for users involved in threads descending from the same post and accounted for 7% of users in our set.

[3]All features except for Facebook age, 28-day activity, average Anger emotion, binary, and slope features are additionally log transformed before standardization.

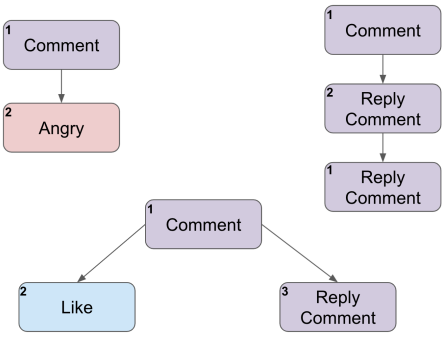


Figure 3: Examples of conversation motifs from a 28-day user motif interval. Each action (e.g., comment) is labeled with a de-identified ID for each user involved in the motif.

them in terms of gender (self-reported), country, age, friend count, Facebook account age, and days of activity on Facebook in the 28 days before the target comment (28-day activity).

To test whether they differed in their involvement in negative online activity (**H2**), we compare users in terms of the conversational motifs they were party to before the target conversation. Motifs are recurrent subgraphs of a larger network graph, showing local patterns within it. The motifs we study are subgraphs of Facebook's conversation graph and encapsulate all conversational interactions [48] under a post. These interactions are made up of comments and reactions, where reactions are icons such as 'angry', 'sad', or 'wow', which allow users to easily express their feelings about a comment or post. Motifs are collected in the form of tuples and triples. We show examples of three different conversational motifs in Figure 3. For example, Motif 1 shows one person giving an angry reaction to another's comment and Motif 2 shows User 2 replying to User 1's comment and then User 1 replying in turn.

We are primarily interested in user motifs, which are the motifs (or interaction patterns) a user was a part of in the 28 days prior to the target conversation. It is important to note that the user IDs for each motif action (e.g., comment) are not related to the user's actual ID on Facebook. For example, Motif 1 in Figure 3 represents an interaction in which a de-identified user labeled "2" gives an angry reaction to User 1's comment. Because of the de-identification, we cannot easily trace a specific action to a specific user and can only determine the type of motif each user was involved in (e.g., in Motif 1, we know the identity of people involved, but not who posted the original comment or the angry reaction). As we are interested in whether conflict users are involved in more negative activity online, we count the number of motifs they are included in that contain one of the seven reaction types: love, like, sorry, support, wow, anger, and haha. Therefore, we are counting instances in which a user has given, received, or interacted with (through a mutual comment) the various reaction types online. We treat the count of each reaction type as a separate variable in the regression model. To test whether conflict users were involved in previous negative online activity (**H2**), we consider the anger reaction as negative and love, like, sorry, and support as positive, with wow and haha as ambiguous.

To test whether conflict commenters are less connected to their group (**H3**), we examine how long a user has been a member of the group and their activity in the group in the 28 days before the target comment, including their number of likes, comments, posts, and reactions.

4.3 Conversation Dynamics

We investigate the relationship between conversational dynamics and conflict in two forms: *average* characteristics and *evolutionary* characteristics. For the former, we calculate the average values of the conversational features (e.g., the average anger during a conversation). When examining the conversation's evolution, we instead calculate how each feature changes during the conversation (e.g., changes in anger from early to late in the conversation). To do so, we split the conversation into three parts (i.e., beginning, middle, and end), where each part contains an equal number of comments, and compute the average feature value for the comments in each time period. We split the conversation into three sections based on the minimum length of the conversation history in our dataset. Since the lengths of the conversations in our dataset vary, the minimum number of comments in a time period is 1 (for short conversations) and the maximum is $max\_length/3$. When examining the inter-comment reply-rate, we split each conversation into two sections (i.e., beginning and end), because there are fewer intervals between comments than there are comments. Using these values, we calculate the best fit line and utilize the slope of the line to represent the feature's evolution throughout the conversation.

When calculating the average and slope values, we started at the top-level comment and ended with the comment before the target comment. We do not include the target comment in the evolution analysis. Studying the average feature values gives us a static comparison of conflict and non-conflict conversations (e.g., Do conflict conversations contain more anger?). Slope values allow us to examine how changes in the variable are related to conflict (e.g., Do conflict conversations become more angry over time?). A slope greater than zero indicates the variable is increasing from the beginning to the end of the conversation, while a slope less than zero indicates it is decreasing.

To test whether conflict conversations become more intense and emotionally negative over time (**H4**), we examine three different features in the conversation: emotions, usage of hate speech, and the speed of the conversation reflected in the inter-reply rate. We utilize two emotion analysis models[4][5], which were both trained on Twitter social media data [31, 35]. We average the results from the two models to account for the variation in the original data collection, where the DistilBERT model [34] was trained on emotion-related hashtags and the RoBERTa model [29] was trained on data collected through an emotion-based keyword list. We consider only emotions measured by both models: anger, joy, and sadness. To do so, we normalize the scores within each model's output for these three emotions to add to one. We then compute the average scores for each emotion across the two models to account for model idiosyncrasies. For each comment in the conversation history, we compute the three emotion scores with this method. For example, the models indicate that C4 in Figure 1a is high in anger. Utilizing the scores for

[4]https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion

each comment, we calculate the average and slope emotion scores for the conversation.

We analyze negativity in the conversation through a different perspective, we examine hate speech usage, which can be defined as offensive or threatening language towards an individual or group. To analyze the usage of hate speech in the conversation, we utilize a hate speech detector that produces a binary classification of whether the text contains hate speech [3][5]. To include a continuous measure of hate speech in our analysis, we use the confidence score from the model. The model indicates that example C4 in Figure 1b, has a higher degree of hate speech than other comments in the conversation, due to the phrase "are you stupid?".

The final variable to operationalize for the intensity of the conversation is the inter-reply rate, with faster replies (i.e. smaller inter-reply rate) indicating a more heated or intense conversation. We compute the time between each consecutive comment pair in the conversation history. When computing the slope of the inter-reply rate, we use the ratio of the average of the second half of the conversation to the first half. This is done to account for some short conversations, which contain only two inter-reply values.

To test **H5** about the group or personal nature of the conflict, consistent with prior research [37, 39], we examine pronoun usage as an indicator of the group (e.g., "we", "us", and "them") or personal nature (e.g., "I" and singular "you") of the conversation. We also examine the number of unique participants in the conversation. We count the number of each type of pronoun in the conversation and how this changes over the course of the conversation with regards to: first, second, and third-person singular and plural pronouns. Because English does not distinguish between second-person singular and plural pronouns, we treat comments with "you" and a user mention as singular (e.g., "User3 ... are you so offended") and those without a user mention as plural. To test whether the conflict conversations contain in-group/out-group qualities (**H5a**), we examine the use of plural pronouns, where an increased usage of plural pronouns can indicate a group identity focus. To test **H5b**, we analyze the usage of singular pronouns and the number of unique commenters in the conversation. To compute the latter, we count the number of unique commenters in the conversation history and divide this by the conversation history length. This ratio represents the number of unique commenters per conversation. For example, if the conversation included two users and contained five comments, the calculated value would be 0.4. An increase in singular pronouns and smaller number of unique users per conversation defines a more personal conversation.

5 RESULTS AND DISCUSSION

The results for the multilevel logistic regression models are shown in Table 2. Given the McFadden's R^2, BIC, and Log Likelihood values, we find that the model fit improves with each set of features added to the model, even when adjusting for its complexity. We provide results for all of our predictor variables, including the user-level controls gender, age, Facebook age, friend count, and 28-day activity. We do not show the results for the country control variables, although they were included in the model. We provide the odds ratio for each feature and discuss the results below.

*Pre-existing User Features (**H1**).* All of the pre-existing user demographic variables except for age are statistically significant. Most interestingly, the results indicate that conflict commenters are more likely to be male, have a newer Facebook account, and fewer friends on Facebook. The user's gender is the strongest predictor, with the odds of a man being involved in a reported conflict about 60% greater than the odds of a woman. This stark contrast is evident in Figure 4a, where the non-conflict commenters were almost two-thirds female, while the conflict commenters were more balanced across genders.

*Involvement in Online Activity (**H2**).* The results for user motifs show conflict commenters are more likely to have a history of engaging in conversations with anger and haha reactions, and less likely to be involved in ones with sorry, support, love, and like reactions. These results suggest that compared to non-conflict commenters, conflict commenters were involved in emotionally negative online conversations and less involved in positive ones (e.g., support). However, because of the de-identified way motifs were collected, these results do not tell us if the conflict commenter was the source of negative reactions, the target of them, or simply a bystander in conversations where these reactions were produced.

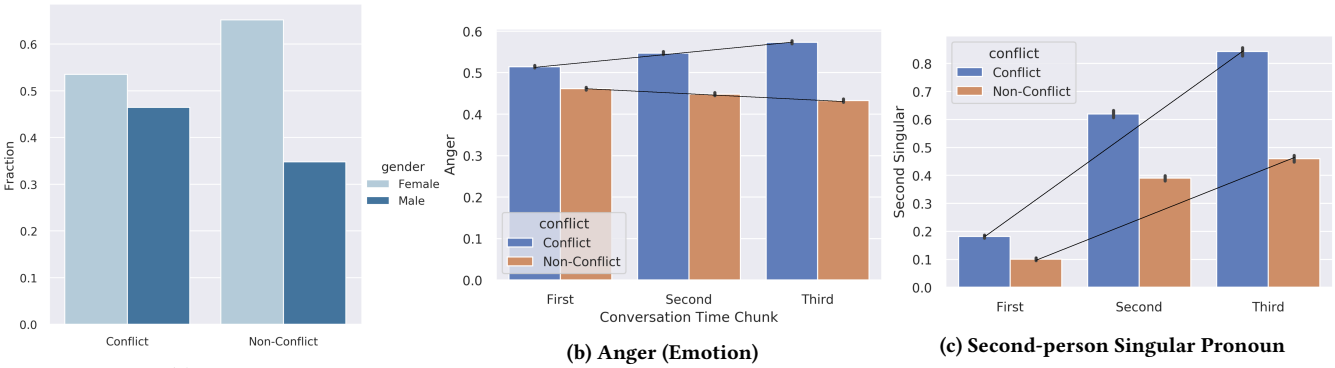
*Connection to Group (**H3**).* Compared to non-conflict commenters, conflict commenters are generally less connected to the group where the conflict occurred. They are less likely to start conversations by initiating posts (OR=0.81) or to positively react with likes to others' posts or comments (OR=0.80). Instead, they are more likely to comment on others' contributions (OR=1.08). These results suggest that conflict commenters are not moving the conversation forward in a positive manner, and are instead reacting negatively towards other members' posts and comments, as we show in the next section that examines the content of their comments. In addition, conflict commenters are members of the group for 16% less time (OR=0.95), with a median of 5.3 months for conflict commenters versus 6.3 months for non-conflict commenters. This is consistent with interpretations that they lacked time to become connected to the group or to understand its norms, or they had joined the group with an intent to create conflict.

*Emotions and Intensity (**H4**).* The average and slope emotion values for the three emotions–anger, sadness, and joy–are highly correlated, with the absolute value of the Pearson correlation ranging from 0.29 to 0.79. To reduce multicollinearity, we include only the anger emotion in the multilevel model. Results show that conflict conversations contain more anger overall (OR=1.61) and that anger increases over the course of the conversation (OR=1.28). The change in anger among conflict and non-conflict conversations is shown in Figure 4b, where conflict conversations start with more anger than non-conflict ones, and the gap increases, with conflict conversations growing angrier while non-conflict become somewhat less angry. This trend is illustrated in Figure 1b, where User 1 first neutrally responds to the post ("Both are average, but I think Spiderman is stronger") but ends in anger ("You're so insecure with yourself...Grow up."). Conflict conversations have a higher overall

[5]The definition for hate speech in this paper is derived from detector described in [3] rather than from Meta's Community Standards.

Table 2: Odds ratios for each variable from the multilevel logistic regression models, where an odds ratio > 1 is positively associated with conflict and odds ratio < 1 is positively associated with non-conflict. We also include p-values for each feature, where *: p<0.05, **: p<0.01, ***: p<0.001.

Feature Type	Feature	H1	H2	H3	H4		H5	
User	Age	0.975*	0.981	0.995	0.990	0.995	1.009	1.011
	Gender (Male)	1.596***	1.360***	1.385***	1.290***	1.380***	1.202***	1.348***
	FB Age	0.891***	0.883***	0.886***	0.903***	0.888***	0.911***	0.890***
	Friend Count	0.926***	0.958**	0.946**	0.944**	0.947***	0.943**	0.947***
	28-day Activity	0.936***	0.941***	0.943***	0.948***	0.942***	0.945***	0.938***
	Love (Motif)		0.897***	0.904***	0.924***	0.903***	0.913***	0.895***
	Like (Motif)		0.922***	0.992	0.988	0.990	0.996	0.997
	Support (Motif)		0.874***	0.890***	0.908***	0.885***	0.903***	0.877***
	Sorry (Motif)		0.867***	0.887***	0.896***	0.891***	0.912***	0.896***
	Wow (Motif)		0.959	0.987	1.005	0.986	1.018	0.993
	Anger (Motif)		1.371***	1.341***	1.267***	1.336***	1.261***	1.338***
	Haha (Motif)		1.243***	1.216***	1.177***	1.223***	1.169***	1.220***
	# of Comments			1.078***	1.086***	1.084***	1.029	1.035
	# of Likes			0.803***	0.802***	0.801***	0.829***	0.821***
	# of Reactions			0.965	0.956*	0.968	0.966	0.978
	# of Posts			0.807***	0.820***	0.806***	0.823***	0.804***
	Group Membership Length			0.953***	0.942***	0.952***	0.949***	0.956***
Conversation					Average	Slope	Average	Slope
	Anger (Emotion)				1.612***	1.280***	1.519***	1.259***
	Hate Speech				1.090***	1.001	1.094***	1.004
	Inter-reply Rate				1.102***	0.944***	1.123***	0.973
	First Singular (Pronoun)						0.871***	1.034
	First Plural (Pronoun)						0.949***	1.002
	Second Singular (Pronoun)						1.536***	1.359***
	Second Plural (Pronoun)						1.060***	1.049***
	Third Singular (Pronoun)						0.984	1.005
	Third Plural (Pronoun)						0.902***	0.930***
	# of Unique Commenters						0.763***	0.737***
	McFadden's R^2	0.016	0.030	0.046	0.086	0.057	0.129	0.087
	BIC	42533.8	42010.8	41372.3	39660.3	40951.5	37906.7	39727.1
	Log Likelihood	-21204.8	-20907.1	-20562.0	-19690.5	-20336.1	-18777.5	-19687.7
	Log Likelihood Ratio (Chisq)		595.4***	690.1***	451.9***	1743.1***	1825.9***	1296.7***



(a) Gender

(b) Anger (Emotion)

(c) Second-person Singular Pronoun

Figure 4: Gender, anger emotion, and second-person singular pronoun differences between conflict and non-conflict commenters and conversations. Figures b and c represent the changes in anger emotion and second-person singular pronoun usage over time throughout the conversation and contain feature slope lines and black bars to reflect 95% confidence intervals.

usage of hate speech (OR=1.09), but this does not increase more (OR=1.00). The results for inter-reply rate show that while conflict conversations tend to have a slower reply rate overall (OR=1.10), replies come more quickly as the conversation progresses (OR=.94). This suggests that while users involved in conflict take longer to explain themselves in the beginning of a conversation, they respond to each other more quickly as the conversation intensifies.

*Personal vs. Group Orientation (**H5**).* To study whether group-oriented language was more common in conflict conversations, we examined the use of first person plural and third person plural pronouns. These were used less in the conflict conversations (OR=.95 and .90 respectively), which is inconsistent with our initial hypothesis and suggests that conflict conversations are less likely than non-conflict ones to involve in-group/out-group talk. Instead, conflict conversations have a much larger usage of second-person singular ("you") pronouns (OR=1.54) and their use increases over the course of the conversation much more rapidly in the conflict conversations (OR=1.36), as shown in Figure 4c. In contrast, conflict conversations have fewer first-person singular ("I" or "me") pronouns (OR=.87) at a rate that doesn't change over the conversation (OR=1.03), suggesting the participants are focused on others and not on themselves [8]. In addition to differences in the use of pronouns reflecting group-oriented versus interpersonal language, conflict conversations involved fewer participants overall (unique commenters OR=.76). Together these results suggest conflict conversations tend to be personal rather than group oriented and their personal nature increases over time. These conclusions are illustrated in Figure 1b, where by the end, the conflict commenter singles out User 1 with personal attacks like "are you stupid" and User 1 retaliates with "You're so insecure".

It is worth noting that the odds ratios differ for some variables when comparing the average and slope columns in H4 and H5, especially for gender and user motifs. This is a statistical artifact due to weak correlations (<0.15 Pearson correlation) between these features and average anger emotion.

6 CONCLUSION

Discussion. In this paper, we analyzed the emergence of online conflicts in group conversations on Facebook. Figure 6 in the appendix shows the results from Table 2's H5 slope model, visualizing the relative importance of each predictor variable. The user findings indicate that some users, for example men, are predisposed towards conflict before the conversation even started. Additionally, conflict commenters are typically not well-integrated into their groups. Our study of conflict evolution shows conflicts become more personal and focused on the other individuals (e.g., increase in singular "you" pronouns) over the course of the conversation and become more heated (e.g., increases in angry language and speed of replies). Together, these results suggest that people in a conflict conversation have weaker ties to the group [40] and a lack of harmony and empathy [8], which may lead to conversations in which people attack each other personally.

Since many of the features examined in our analysis are not specific to Facebook, our methodology can be generalized to non-Facebook related environments (e.g., Reddit & Twitter). Specifically, the analysis of conversational dynamics, pre-existing (H1) and group connection (H3) features are available on other platforms as well. Although our study investigates only online conflicts, we hypothesize that specific findings, such as the importance of change in anger within the conversation, may generalize across online and offline conversations. To the extent to which the nature and evolution of conflicts in offline and online groups are similar, because of the fine-grained, archival nature of the data, the study of online conflict may lead to insights that cannot be gained from studying offline ones. However, online conflicts may differ from offline ones because of the large size and fluid nature of online groups, the likelihood of communication among strangers, and the relatively public nature of online group communication.

The study of online conflicts is also practically important, in guiding the improvement of social media technologies to better adapt to the unique aspects of online interactions. This type of research could lead to the creation of better conflict detection tools to identify conflicts as they emerge rather than after they have already occurred. Such tools can enable administrators and moderators of groups to be notified of an emerging conflict earlier and to get involved to mediate the conversation before it erupts or to warn participants in emerging conflicts to cool down.[6] Early conflict detection, by examining participants' pre-conversation behavior and early conversational dynamics, can also enable automated notifications for users involved in the conversation, warning them before responding to an escalating conflict and reminding them to respond empathetically.

Limitations. A limitation when studying pronoun usage is the various ways to express specific pronouns, such as "ya" for "you" and "y'all" to indicate a plural version of "you". Instead, we only study the Standard American English versions of these pronouns. Additionally, our decision to create paired conflict/non-conflict samples to control for the post's and group's content prevents us from examining differences in the topic of these conversations. Despite the matching, there is a possibility that topics in the paired conversation threads may drift, introducing content-based influences.

Next Steps. Some of the limitations described above can serve as a stepping stone for future research. For example, one can move beyond exclusively relying on pronouns to identify identity-based conversations, and use other metrics such as demographic-type keywords within the conversation [41]. Additionally, studying the specific thread's topic can aid in recognizing whether conversations evolve to a different topic over time [2]. Although we examined the number of participants and how language changed over the course of a conversation, we have not examined other aspects of group dynamics, such as new participants joining the conversation after it has started and the influence of coalition formation, e.g., [25]. Furthermore, we examine negative emotions and hate speech, but not the influence of positive types of conversational language, such as politeness and agreeability [4, 45]. Perhaps what is needed most is a richer theoretical understanding of the social psychology of conflict, such as exploratory work by Weingart et al. [44], which can serve as a framework on which to scaffold the types of empirical understanding available from studying online conflict.

[6] See https://www.theverge.com/2021/10/6/22713211/twitter-pre-tweet-prompt-fight-intense-conversation.

REFERENCES

[1] Mahmoud Al-Ayyoub, Abdullateef Rabab'ah, Yaser Jararweh, Mohammed N Al-Kabi, and Brij B Gupta. 2018. Studying the controversy in online crowds' interactions. *Applied Soft Computing* 66 (2018), 557–563.

[2] Hind Almerekhi, Haewoon Kwak, Joni Salminen, and Bernard J Jansen. 2020. Are these comments triggering? Predicting triggers of toxicity in online discussions. In *Proceedings of The Web Conference 2020*. 3033–3040.

[3] Sai Saket Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep Learning Models for Multilingual Hate Speech Detection. *arXiv preprint arXiv:2004.06465* (2020).

[4] Jiajun Bao, Junjie Wu, Yiming Zhang, Eshwar Chandrasekharan, and David Jurgens. 2021. Conversations Gone Alright: Quantifying and Predicting Prosocial Outcomes in Online Conversations. In *Proceedings of the Web Conference 2021*. 1134–1145.

[5] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 1644–1650.

[6] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial behavior in online discussion communities. In *Ninth International AAAI Conference on Web and Social Media*.

[7] Minje Choi, Luca Maria Aiello, Krisztián Zsolt Varga, and Daniele Quercia. 2020. Ten social dimensions of conversations and relationships. In *Proceedings of The Web Conference 2020*. 1514–1525.

[8] Cindy Chung and James W Pennebaker. 2007. The psychological functions of function words. *Social communication* 1 (2007), 343–359.

[9] Mauro Coletto, Kiran Garimella, Aristides Gionis, and Claudio Lucchese. 2017. Automatic controversy detection in social media: A content-independent motif-based approach. *Online Social Networks and Media* 3 (2017), 22–31.

[10] CKW de Dreu. 2010. Social conflict: the emergence and consequences of struggle and negotiation. *Handbook of Social Psychology: Vol. 2* (2010), 983–1023.

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

[12] 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)*.

[13] Frank RC De Wit, Lindred L Greer, and Karen A Jehn. 2012. The paradox of intragroup conflict: a meta-analysis. *Journal of applied psychology* 97, 2 (2012), 360.

[14] Frank R.C. de Wit, Karen A. Jehn, and Daan Scheepers. 2013. Task conflict, information processing, and decision-making: The damaging effect of relationship conflict. *Organizational Behavior and Human Decision Processes* 122, 2 (2013), 177–189. https://doi.org/10.1016/j.obhdp.2013.07.002

[15] Shiri Dori-Hacohen and James Allan. 2013. Detecting controversy on the web. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 1845–1848.

[16] Shiri Dori-Hacohen and James Allan. 2015. Automated controversy detection on the web. In *European Conference on Information Retrieval*. Springer, 423–434.

[17] Alice H Eagly and Valerie J Steffen. 1986. Gender and aggressive behavior: a meta-analytic review of the social psychological literature. *Psychological bulletin* 100, 3 (1986), 309.

[18] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing* 1, 1 (2018), 1–27.

[19] Harold Guetzkow and John Gyr. 1954. An analysis of conflict in decision-making groups. *Human relations* 7, 3 (1954), 367–382.

[20] Richard M Guo. 2008. Stranger danger and the online social network. *Berkeley Technology Law Journal* 23, 1 (2008), 617–644.

[21] Jack Hessel and Lillian Lee. 2019. Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features. In *Proceedings of NAACL-HLT*. 1648–1659.

[22] Michael A Hogg. 2020. *Social identity theory*. Stanford University Press.

[23] Jennifer L Holt and Cynthia James DeVore. 2005. Culture, gender, organizational role, and styles of conflict resolution: A meta-analysis. *International Journal of Intercultural Relations* 29, 2 (2005), 165–196.

[24] Karen A Jehn. 1997. A qualitative analysis of conflict types and dimensions in organizational groups. *Administrative science quarterly* (1997), 530–557.

[25] Karen A Jehn and Katerina Bezrukova. 2010. The faultline activation process and the effects of activated faultlines on coalition formation, conflict, and group outcomes. *Organizational Behavior and Human Decision Processes* 112, 1 (2010), 24–42.

[26] Kathleen A Kennedy and Emily Pronin. 2008. When disagreement gets ugly: Perceptions of bias and the escalation of conflict. *Personality and Social Psychology Bulletin* 34, 6 (2008), 833–848.

[27] Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community interaction and conflict on the web. In *Proceedings of the 2018 world wide web conference*. 933–943.

[28] Giuseppe Labianca, Daniel J Brass, and Barbara Gray. 1998. Social networks and perceptions of intergroup conflict: The role of negative relationships and third parties. *Academy of Management journal* 41, 1 (1998), 55–67.

[29] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[30] Yelena Mejova, Amy X Zhang, Nicholas Diakopoulos, and Carlos Castillo. 2014. Controversy and sentiment in online news. *arXiv preprint arXiv:1409.8152* (2014).

[31] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*. 1–17.

[32] Thomas F Pettigrew, Linda R Tropp, Ulrich Wagner, and Oliver Christ. 2011. Recent advances in intergroup contact theory. *International Journal of Intercultural Relations* 35, 3 (2011), 271–280.

[33] Lee Ross and Andrew Ward. 1995. Psychological barriers to dispute resolution. In *Advances in experimental social psychology*. Vol. 27. Elsevier, 255–304.

[34] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[35] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized Affect Representations for Emotion Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3687–3697. https://doi.org/10.18653/v1/D18-1404

[36] Martin Saveski, Brandon Roy, and Deb Roy. 2021. The Structure of Toxic Conversations on Twitter. In *Proceedings of the Web Conference 2021*. 1086–1097.

[37] Andreas Spitz, Ahmad Abu-Akel, and Robert West. 2021. Interventions for Softening Can Lead to Hardening of Opinions: Evidence from a Randomized Controlled Trial. In *Proceedings of the Web Conference 2021*. 1098–1109.

[38] Henri Tajfel. 1974. Social identity and intergroup behaviour. *Social Science Information/sur les sciences sociales* 13, 2 (1974), 65–93. http://www.sagepublications.com/

[39] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*. 613–624.

[40] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[41] Anna Tigunova, Andrew Yates, Paramita Mirza, and Gerhard Weikum. 2019. Listening between the lines: Learning personal attributes from conversations. In *The World Wide Web Conference*. 1818–1828.

[42] Benjamin Timmermans, Tobias Kuhn, Kaspar Beelen, and Lora Aroyo. 2017. Computational controversy. In *International Conference on Social Informatics*. Springer, 288–300.

[43] Harry C Triandis. 2000. Culture and conflict. *International journal of psychology* 35, 2 (2000), 145–152.

[44] Laurie R Weingart, Kristin J Behfar, Corinne Bendersky, Gergana Todorova, and Karen A Jehn. 2015. The directness and oppositional intensity of conflict expression. *Academy of Management Review* 40, 2 (2015), 235–262.

[45] Michael Yeomans, Alejandro Kantor, and Dustin Tingley. 2018. The politeness Package: Detecting Politeness in Natural Language. *R Journal* 10, 2 (2018).

[46] Michael Yeomans, Julia Minson, Hanne Collins, Frances Chen, and Francesca Gino. 2020. Conversational receptiveness: Improving engagement with opposing views. *Organizational Behavior and Human Decision Processes* 160 (2020), 131–148.

[47] 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)*. Association for Computational Linguistics, Melbourne, Australia, 1350–1361. https://doi.org/10.18653/v1/P18-1125

[48] Justine Zhang, Cristian Danescu-Niculescu-Mizil, Christina Sauper, and Sean J Taylor. 2018. Characterizing online public discussions through patterns of participant interactions. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–27.

A APPENDIX

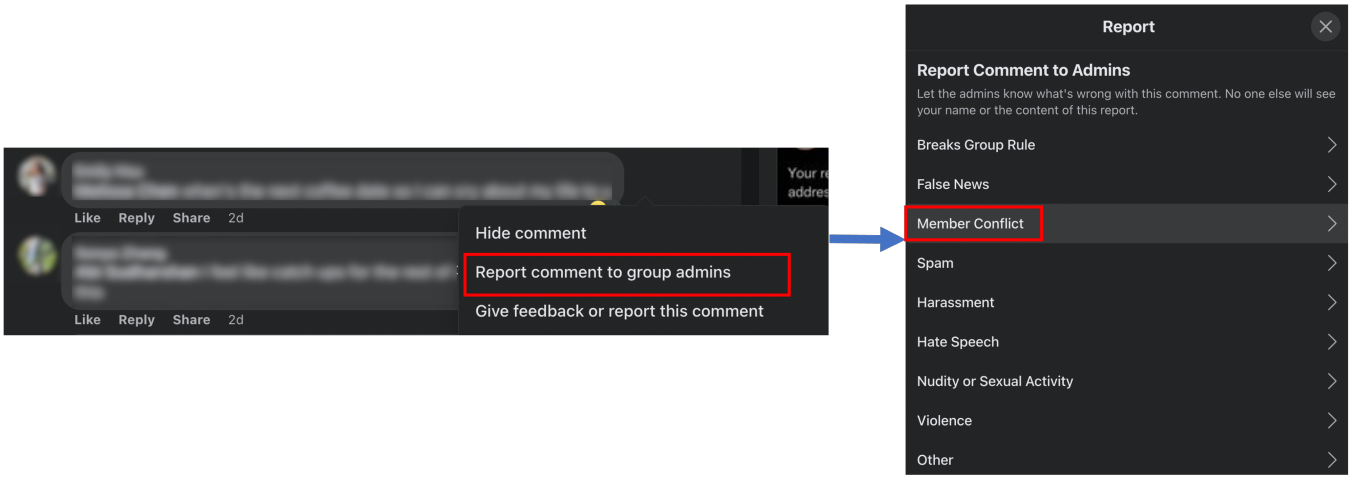


Figure 5: Visualization of the conflict comment reporting tool available within Facebook groups. Only group members are allowed to report comments to the group admins.

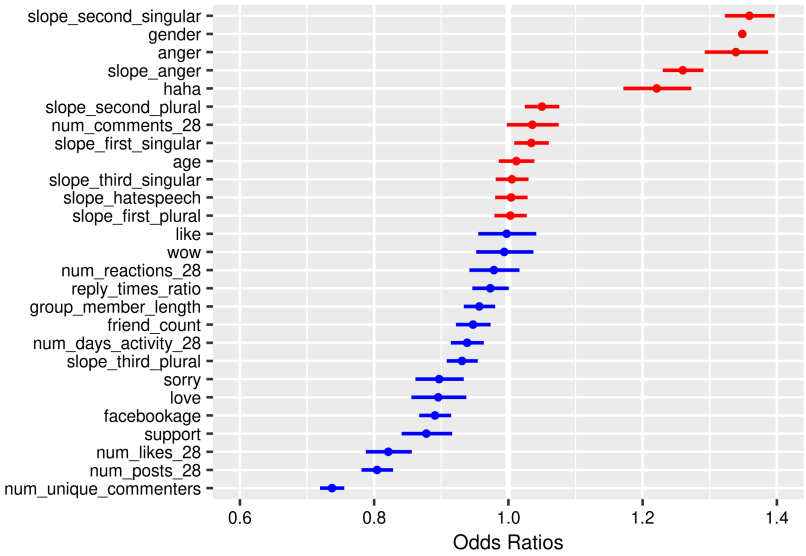


Figure 6: Visualization of odds ratios for each feature from the H5 slope model, where an odds ratio > 1 is positively associated with conflict and odds ratio < 1 is positively associated with non-conflict.

Table 3: Mean, median, and standard deviation for each feature in the model before data transformation and the comment/conversation lengths. Statistics are separated by conflict and non-conflict samples.

Variable	Mean		Median		Standard Deviation	
	Conflict	Non-Conflict	Conflict	Non-Conflict	Conflict	Non-Conflict
Age	41.570	42.122	39.000	40.000	15.066	15.2481
Gender(% Male)	46.45	34.81	N/A	N/A	N/A	N/A
FB Age(months)	118.457	127.265	141.900	147.233	54.222	50.237
Friend Count	627.606	659.820	355.0	403.0	817.862	814.331
28-day Activity	27.452	27.672	28.000	28.000	2.517	1.868
Love (Motif)	467.0394	520.191	97.000	127.000	4188.598	2127.702
Like (Motif)	7389.856	8328.671	1344.0	1409.0	80283.648	85743.442
Support (Motif)	71.765	92.203	15.0	20.0	697.769	458.485
Sorry (Motif)	43.966	49.629	6.000	8.000	550.891	291.575
Wow (Motif)	48.268	52.701	12.0	13.0	375.994	196.294
Anger (Motif)	59.47	50.623	4.000	3.000	675.640	559.039
Haha (Motif)	1812.686	1729.279	269.0	245.0	19643.840	11898.444
# of Comments	23.556	34.374	7.000	9.000	62.826	95.836
# of Likes	19.351	32.792	6.000	3.000	73.393	108.912
# of Reactions	11.574	20.167	2.000	3.000	45.971	78.209
# of Posts	0.709	1.993	0.000	0.000	4.728	22.983
Group Membership Length(months)	12.656	14.166	5.332	6.303	18.182	19.262
Anger Average (Emotion)	0.540	0.436	0.542	0.428	0.181	0.209
Anger Slope (Emotion)	0.0287	-0.016	0.026	-0.0138	0.181	0.181
Hate Speech Average	0.108	0.092	0.078	0.064	0.086	0.079
Hate Speech Slope	0.002	0.000	0.000	0.000	0.087	0.080
Inter-reply Rate Average (hours)	1.32	1.12	0.35	0.31	2.51	2.20
Inter-reply Rate Ratio	3.841	5.332	0.468	0.611	33.068	31.845
First Singular Average(Pronoun)	0.640	0.718	0.428	0.500	0.723	0.840
First Singular Slope(Pronoun)	0.113	0.050	0.000	0.000	0.755	0.820
First Plural Average(Pronoun)	0.127	0.146	0.000	0.000	0.279	0.326
First Plural Slope(Pronoun)	-0.008	-0.009	0.000	0.000	0.304	0.343
Second Singular Average(Pronoun)	0.494	0.279	0.333	0.000	0.552	0.423
Second Singular Slope(Pronoun)	0.322	0.163	0.000	0.000	0.607	0.456
Second Plural Average(Pronoun)	0.164	0.150	0.000	0.000	0.325	0.328
Second Plural Slope(Pronoun)	-0.144	-0.162	0.000	0.000	0.399	0.436
Third Singular Average(Pronoun)	0.260	0.283	0.000	0.000	0.572	0.655
Third Singular Slope(Pronoun)	-0.010	-0.023	0.000	0.000	0.528	0.572
Third Plural Average(Pronoun)	0.246	0.273	0.000	0.000	0.411	0.464
Third Plural Slope(Pronoun)	-0.028	-0.009	0.000	0.000	0.423	0.461
# Unique Commenters/Conversation	0.521	0.576	0.500	0.500	0.181	0.173
Comment Length (words)	31.636	28.588	21.000	19.000	32.645	30.691
Conversation Length (comments)	5.864	4.844	5.000	4.000	2.492	1.618