Conversations Gone Alright: Quantifying and Predicting Prosocial Outcomes in Online Conversations

Jiajun Bao*
jiajunb@andrew.cmu.edu
Carnegie Mellon University, Language Technologies Institute
USA

Junjie Wu*
junjie.wu@connect.ust.hk
The Hong Kong University of Science and Technology
China

Yiming Zhang*
yimingz@umich.edu
University of Michigan
USA

Eshwar Chandrasekharan
eshwar@illinois.edu
University of Illinois, Urbana-Champaign
USA

David Jurgens
jurgens@umich.edu
University of Michigan, School of Information
USA

ABSTRACT

Online conversations can go in many directions: some turn out poorly due to antisocial behavior, while others turn out positively to the benefit of all. Research on improving online spaces has focused primarily on detecting and reducing antisocial behavior. Yet we know little about positive outcomes in online conversations and how to increase them—is a prosocial outcome simply the lack of antisocial behavior or something more? Here, we examine how conversational features lead to prosocial outcomes within online discussions. We introduce a series of new theory-inspired metrics to define prosocial outcomes such as mentoring and esteem enhancement. Using a corpus of 26M Reddit conversations, we show that these outcomes can be forecasted from the initial comment of an online conversation, with the best model providing a relative 24% improvement over human forecasting performance at ranking conversations for predicted outcome. Our results indicate that platforms can use these early cues in their algorithmic ranking of early conversations to prioritize better outcomes.

CCS CONCEPTS

• **Human-centered computing**; • **Social and professional topics**; • **Applied computing** → *Sociology*;

KEYWORDS

prosocial behavior, antisocial behavior, social media, behavioral forecasting

ACM Reference Format:
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 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, **??** pages. https://doi.org/10.1145/3442381.3450122

*Authors contributed equally to this research.

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3450122

Post: "Studied like crazy all last week and finally took a Technician course and tested on Saturday. Boom, I'm legal! Thanks for all of the support I've got on /r/amateurradio!!"
⌞ **Reply 1**: That test is a killer!
⌞ **Reply 2**: Great job! Hope to hear you on the air some day

Figure 1: Which reply is likely to lead to a positive, productive conversation? Here, we introduce new metrics for measuring the prosocial qualities of social media discussions and develop new models to predict which of these replies will lead to a conversation with higher prosocial behavior.

1 INTRODUCTION

Interacting with others online has become a common facet of daily life. Yet, these interactions often can turn out poorly, in part due to toxic behavior on the part of others [28, 48]. Given the significant impact of experiencing these negative activities on well-being [34, 66, 67], substantial research effort has been put into detecting such toxic behavior and facilitating platform tools to remove it. However, despite sophisticated techniques for measuring antisocial behavior, the key metrics for *prosocial* behavior are relatively unknown, to the point that major platforms such as Twitter and Instagram have both called for researchers to develop such metrics [30, 74]. Here, we operationalize theories from social psychology to quantify and measure prosocial behavior in social media, showing a rich diversity in the types of behaviors, and then show that these positive outcomes can be forecasted from the early stages of a conversation.

The impact of online discussions on daily life and mental health has prompted multiple studies on online conversational dynamics. A significant effort has focused on detecting antisocial behaviors such as hate speech [79] , trolling [21] , or bullying [52], in attempts to mitigate their effect by offering moderators tools to find and remove them. Further, recent work has shown that the initial start of a conversation can forecast antisocial outcomes from early linguistic and behavioral information [19, 39, 52, 85]. However, only a handful of studies have examined prosocial behaviors, such as making constructive comments [45, 46, 59] or offering supportive

messages [61, 77, 78]. Our work brings together these lines of research through a systematic examination of prosocial behaviors and building models to forecast these conversational outcomes.

We introduce eight types of prosocial metrics and develop new methods to forecast the prosocial trajectory of a conversation from its early interactions. Focusing this task on one of prediction is strongly motivated by the implications for real-world impact. Online platforms regularly engage in content reranking where comments and threads are reordered according to internal objectives [14, 51]. Given the ability to predict the prosocial trajectory of a conversation, platforms can potentially rerank the initial comments to a post (or other comments) to emphasize those that will lead to better community experiences.

This paper offers the following three contributions. Using theoretical insights from prior literature on prosocial and antisocial behavior in online and offline contexts (§2), we introduce a panel of prosocial metrics and construct a large-scale corpus of social media conversations labeled by these outcomes (§3). Using this corpus, we demonstrate that these metrics are significantly associated with human judgments of prosociality and show that prosociality is not just the absence of antisocial behavior. Second, we introduce four new models for forecasting the prosocial quality of a conversation (§5), showing that such outcomes can be accurately forecasted from cues early in the conversation. Third, given the first comment of two conversations, we demonstrate that both our models and people can forecast which conversation is likely to turn out better (§6), with our model offering a 24% improvement on human accuracy with respect to chance. Our work has implications for platforms' abilities to surface online interactions in order to create positive outcomes for individuals participating in them.

2 PROSOCIAL BEHAVIOR

Prosocial behavior began as an antonym used by social scientists for describing the opposite of antisocial behavior [6, 44], with Mussen and Eisenberg-Berg defining the behavior as "voluntary actions that are intended to help or benefit another individual or group of individuals." Since this time, prosocial behavior has broadened to include a range of activities: helping, sharing, comforting, rescuing, and cooperating [9]. Our work examines prosocial behavior in online discussions by deriving a large cohort of candidate metrics for measuring conversations from theory and then testing which are associated with judgments of prosocial behavior online.

The concept of prosociality is complex and the nuances of which aspects of behavior contribute to its perception online are not yet well understood. A few recent approaches examining specific factors related to positive conversational outcomes like constructive comments [59, 60], politeness [24], supportiveness [78], or empathy [15, 70, 86]; or, showing that, in general, online prosocial behaviors mirror offline trends [82]. In the majority of cases, only individual dimensions have been analyzed; however, we note that recent work has proposed studying these dimensions jointly in relationships and social interactions [22] using the ten social dimensions outlined in Deri et al. [25]. Similar to the present work, Choi et al. [22] examines general factors from sociological and psychological literature for relationships to study interactions; however, the factors used here are specifically grounded in prosocial literature and include behavioral factors in addition to linguistic factors. A few studies have tried to measure prosocial behavior as a single variable [32, 33]; however, these approaches in practice have used lexicons that recognize only a subset of the possible prosocial behaviors focused on collective interest and interpersonal harmony.

Prosocial behaviors can take many forms depending on the parties involved and their needs. Here, we identify eight broad categories of behavior ground in prior work from Social Psychology that can be easily operationalized using NLP techniques and are markers of direct prosocial behavior or are behaviors that serve as a precursor to prosocial behavior. As many of these behaviors have been identified and studied in offline settings, our aim is to study how these behaviors are interpreted in online settings in order to curate a set of prosocial metrics that match peoples' readings of online interactions. Following, we outline each category, its connection to prior theory, and summarize how its behavior is recognized. Additional details for metrics and classifiers are provided in Appendices A.1–A.8.

Information Sharing. Individuals seek out information online where others may provide suggestions. In some settings these efforts are codified around collaboratively creating information goods like Wikipedia or open source projects [71]. In social media like Reddit, responses to questions create persistent knowledge that can be learned from by others. This knowledge transfer may take the form of explicit information or references to websites such as Wikipedia or StackOverflow. Here, we operationalize these information sharing behaviors in two ways. First, using information-providing based subreddits (e.g., `r/AskScience`), we train a classifier to recognize informative replies to questions; and then as a prosocial metric, use the classifier to identify and count these replies in a conversation—i.e., does a discussion lead to information sharing? Second, recognizing that URLs often serve as important sources of third-party information, we include counts for (i) information-based domains, e.g., wikipedia.org, and (ii) for all other websites, recognizing that many domains may serve informative purposes (e.g., linking to a product review). More details about the training of the classifier can be found in Appendix A.1.

Gratitude. Gratitude serves an important function in fostering social relationships and promoting future reciprocity [8, 29]. Gratitude not only reinforces existing prosocial behavior, but also motivates more prosocial behavior and itself serves as an indicator that prosocial behavior has occurred [2, 55, 56]. Here, we identify gratitude through a fixed lexicon of phrases that signal gratitude (Appendix A.4), e.g., "thank you," and count how many times such phrases are used in a discussion.

Esteem Enhancement. Prosocial behavior is known to be motivated by a person's self-esteem [9]. Individuals may seek out opportunities to behave prosocially as a way to repair or improve their self-esteem [12]; improved esteem can increase the perception of reciprocity for help, further motivating prosocial actions. Thus, esteem-enhancing actions can serve as a useful behavior to monitor as precursors for prosocial behavior. Here, we measure esteem enhancement using three metrics. First, recognizing that politeness is often used to signal social status [13], we use a politeness

regressor based in part on Danescu-Niculescu-Mizil et al. [24] to measure the average politeness of comment interactions (Appendix A.7); the underlying hypothesis is that more polite messages increase the status (esteem) perception of the recipient. Second, we identify all statements with second-person pronominal references (e.g., you) and count how many have strongly positive sentiment, which approximates identifying compliments (Appendix A.5). Third, we measure the total score given to responses in a discussion as a measure of esteem given by the community to the conversation. The score of a comment is closely correlated with the number of upvotes it receives (as derived through a proprietary measure) and receiving upvotes is known to be an esteem enhancing action [16].

Social Support. In times of distress, individuals turn to their social network for support [9, 80]. Online communities and platforms have provided a parallel support mechanism around many types of needs, such as physical and mental health [11, 57, 77] and weight-loss. Moreover, beyond specific needs, individuals offer supportive messages in general on these platforms, e.g., encouragement [78]. Due to the diversity of support expected on Reddit, we develop a computational model to recognize supportive messages using the data of Wang and Jurgens (Appendix A.8) and use the average supportiveness of comments in a conversation.

Social Cohesion. Social ties create a sense of community, which carries with it the benefits of group membership and altruistic behavior between members [1, 35]. Conversely, exclusions from a group or a weakening of ties decrease prosocial behavior [73]. Indeed, Prinstein and Cillessen note that helping someone join a conversation is a core prosocial behavior and studies have shown that increased linguistic accommodation [62, 72] is associated with increased prosocial behavior [47] and trust [69].

To measure the formation of social bonds, we use four categories of metrics. First, building on the insight of Kulesza et al. [47], we measure linguistic accommodation between commenters using the methods of Danescu-Niculescu-Mizil et al.. Second, recognizing that increased conversation gives rise to social ties, we include metrics for (i) the total number of participants in a conversation, (ii) the longest number of sustained turns between two people, and (iii) the depth of the conversation's comment tree. Third, laughter, as a function of humor, is known to create positive affect between peers and increase cohesion among group members [4, 36, 64]. Therefore, we count the number of laughter events in a conversation (see Appendix A.2 for details). Fourth, self disclosure is known to strengthen social ties [27, 41, 42]; to measure disclosures, we follow prior work [5, 7, 40] and count the number of comments including a first person pronoun.

Fundraising and Donating. An offline prosocial behavior that readily transfers to online behavior is fundraising for charitable activities [71, 83]. Many online charities have websites set up to receive donations and sites such as gofundme for more individualized fundraising activities. Here, we count the number of URLs to these types of sites using a list of popular charities, detailed in Appendix A.6.

Mentoring. Individuals can give expertise through mentoring or advice when others are having a problem, which is a known prosocial behavior [65, 83]. To recognize advice giving and mentoring, we train a classifier to distinguish the language of advice in the responses of advice-based subreddits (e.g., `/r/FashionAdvice`, `/r/RelationshipAdvice`, `/r/LegalAdvice`) from responses in other subreddits. Then, we use this classifier to recognize and count the number of advice-based responses in a conversation thread. Further details on the classifier can be found at A.3.

Absence of Antisocial Behavior. Is prosocial behavior implied by the absence of antisocial behavior? If true, this question suggests that platforms' efforts to find and remove antisocial behavior are sufficient for retaining and promoting prosocial behavior. To answer this question, we label all replies using the Perspective API [84], which assigns a score for toxicity. Highly toxic content like explicit hate speech is assigned scores closer to 1, while more casually offensive language is typically scored closer to the positive decision boundary of 0.5. Here, we consider two definitions for antisocial comments: if the comment's score is above the standard decision boundary (0.5) or a higher boundary (0.8) for highly toxic content. As metrics, we include both (i) the count of comments exceeding both threshold as well as (ii) the percentage of non-toxic comments; the latter percentage allows us to model large discussions where some sub-discussions turn antisocial, but the majority of content is not antisocial.

We note that toxicity itself can be challenging to measure [75]. Multiple models have been proposed for handling different aspects of antisocial behavior [31] and that these models frequently have gaps in what they recognize [3, 43], can encode biases with respect to social groups [68], and be susceptible to adversarial attacks [37]. Nevertheless, as a widely-used measure, Perspective API provides a replicable—though imperfect—tool specifically designed for recognizing multiple forms of toxicity in online conversations in comments, which is the medium studied here.

3 PROBLEM DEFINITION AND DATA

This study's goal is to forecast the future behavior of a conversation from early signals. Specifically, we aim to predict whether the initial comment in a conversation will signal eventual prosocial behavior. This forecasting goal mirrors analogous work in antisocial behavior on conversational trajectories [85], as a first step towards understanding prosocial evolution in conversations.[1] We pursue this goal using data from Reddit, a large social media platform where users create posts and participate in threaded conversations relating to that post. Critically, Reddit provides millions of conversations across different communities, known as subreddits, which span a variety of topics and interaction styles.

To analyze conversations, we extract all conversational threads under a Top-level Comment (Tlc) to a post; such comments are typically made in response to the post itself and serve as conversation starters for the rest of the community. We filter these conversations by removing those where the Tlc (i) has more than 3500 words, as manual inspection showed these were frequently spam

[1] We also acknowledge that other setups have used more conversation as context for forecasting, e.g., Chang and Danescu-Niculescu-Mizil [19] and Liu et al. [52]. While these too are valid setups, we focus on the initial comment intentionally to see whether emerging prosocial conversations can be quickly identified and prioritized.

Prosocial Metric	Description	MCC	Percentage
Information sharing	number of replies classified as informative	0.3452***	0.1218
Link replies	number of links (urls) in all replies	0.5129***	0.1095
Educational Link replies	number of educational links (urls) in all replies	0.6237***	0.0070
Gratitude	number of gratitude in all replies	0.3154***	0.1096
Politeness	average politeness value of all replies	0.0346	1.0000
Linguistic accommodation	Replies to the TLC mirror function word usage	0.0997†	0.5208
Community score	numeric sum of every reply's score	0.5531***	1.0000
Supportiveness	average supportiveness value of all replies	0.0511	1.0000
Subsequent comments	a top comment's total number of replies	0.5988***	1.0000
Direct replies	number of replies responding directly to the TLC	0.1529***	1.0000
Conversation depth	length of the longest replies' thread	0.5400***	1.0000
Sustained conversation partners	number of distinct user pairs appear in all replies	0.5508***	0.4351
Sustained conversations	number of turns in the longest two-person conversation	0.6212***	0.4351
Compliments	number of compliments	0.5374***	0.0045
Laughter	number of expressions of laughter	0.2678***	0.0716
Personal disclosure	number of statements an authors makes about themself	0.4091***	0.3302
Donations	number of links to charities and donation sites	0.2987	0.0002
Mentorship	number replies classified as mentoring	0.3957***	0.0950
% of non-toxicity (untuned)	percentage of non toxic replies in all replies	0.3363***	0.1392
% of non-toxicity (tuned)	percentage of non extremely toxic replies in all replies	0.1222	0.0239
Untuned toxic language	number of toxic replies	−0.2852***	0.1392
Tuned toxic language	number of extremely toxic replies	−0.1014	0.0239

Table 1: Theory-based metrics used to measure prosocial outcomes in conversations (colored by category) and their correlation with human judgments of prosociality, using Matthew's Correlation Coefficient (MCC; see §4 for details) and their rates of occurrence in conversations. Throughout the paper, we use * to denote p<0.001, ** p<0.01, and * p<0.05, here shown after Bonferroni correction. †Acommodation varied in significance across all types, with 10 having a significant MCC (shown here with the mean).**

	Train	Dev	Test
TLC	4,290,361	10,844,404	10,844,281
Subreddits	11,992	53,675	53,650

Table 2: Dataset sizes; note that training data has been downsampled for computational tractability.

posts, or (ii) has been deleted by the user or removed by a moderator. Additionally, Reddit includes a small number of bot accounts that interact in these conversations (e.g., replying with the number of "ooofs" a user has made); to avoid any confounding effect of these bots, we remove all threads containing a comment by a bot account, drawn from a known list of bots[2] and a manually-curated list based on inspection of high-frequency-posting accounts.

The final dataset was constructed from a single year of Reddit activity using the ~500M comments posted between January 2017 and December 2017. This process yielded 140.5M total conversations (unique TLC) across 66K subreddits randomly partitioned into training, development, and test sets using an 8:1:1 ratio (Table 2). Due to computational limitations in labeling each conversation with all metrics (e.g., scoring all comments using the rate-limited Perspective API), we downsample the training data to 4.3M conversations, requiring all subreddits to have at least 100 conversations and keeping at most 500 conversations per subreddit.

4 JUDGING CONVERSATION PROSOCIALITY

What types of conversations do people find more prosocial? The multifaceted nature of prosociality makes numerically rating conversations challenging. Therefore, we answer this question by framing the rating as a paired choice task: given two conversations, select which conversation contains more prosocial behavior.

This binary rating setup also allows us to directly evaluate whether each proposed prosocial metric aligns with human judgments. For each metric (§2), we measure the strength of association with human judgments by computing Matthews Correlation Coefficient (MCC) from a 2x2 contingency table of which conversation in the pair had a higher metric score versus which was selected by annotators as being more prosocial.[3]

Annotated data was collected in two phases. An initial 2000 conversation pairs were collected by sampling two TLC made to the same post using two strategies: (1) 1000 pairs of conversations made any time after the post was authored and (2) 1000 pairs made

[2]https://www.reddit.com/r/autowikibot/wiki/redditbots

[3]While related to the χ^2 measure, MCC differs in that it measure the *strength* of association, rather than just whether the difference in the ratings is statistically significant.

within 5 minutes of their post. After this initial selection, an additional 388 pairs of conversations were included to ensure that each metric occurred in at least 100 pairs, with the exception of the Donation metric, which only occurred in 78 pairs. Two annotators participated in three rounds of training and then divided up the annotations. Annotators attained a Krippendorff's α of 0.78 on 300 mutually-labeled pairs, indicating high agreement.

Results. Table 1 summarizes all of the metrics and their correlations, revealing that most of the metrics predicted from theory are significantly associated with human judgments of prosociality. Four trends merit noting. First, metrics for the breadth and depth of conversations were most correlated with prosocial judgments, with the strongest association for sustained conversation between people; these behavior promote social cohesion and, given the discussion-focused nature of Reddit, are easily measured in conversation. Second, the second-most associated category was for information-providing behaviors. While less common in Reddit conversations as a whole, these actions help other users meet their information needs. Third, surprisingly, prior metrics suggested for prosocial behaviors, politeness and supportiveness, were positive but not significantly associated. We view this negative result as requiring further investigation to confirm, as more precise models for measuring these behaviors may provide more insight. Fourth, the metric around toxicity showed that, indeed, the absence of toxicity was only moderately correlated with human judgments of prosocial conversations, while other types of behavior were more associated. Further, the presence of extremely toxic language, though rare, was not correlated, indicating that a broader picture of toxic behavior—not just extreme events—is necessary for prosocial judgment. This result also indicates that for platforms to measure their health, new metrics like those proposed above are needed and not simply measure the absence of antisocial behavior. However, as our adopted measure of toxicity is a coarse-grain estimate, a potential avenue for future work is to examine whether the absence of specific forms of antisocial behavior or toxicity might individually be associated with the perception of prosocial behavior.

Synthesizing Prosociality. Prosocial behaviors often share a common motivation as individuals seek to engage constructively with one another. Thus, with their conceptual and thematic similarities, a single conversation may contain several of these prosocial behaviors. Given their shared motivation, we ask whether the prosocial metrics could be summarized with a single proxy metric? To test this, we adopt the approach of [76] for synthesizing a single metric from related prosocial behaviors around respect and compute the Principal Component Analysis (PCA) over all the metric scores for conversations in the training data. PCA decomposes these in a set of underlying latent behaviors, capturing the inherent correlations between metrics.

As shown in Figure 2, the first PCA component explains 57.4% of the variance in the prosocial metrics' values and positively loads on all prosocial metrics (Appendix Figure 7), suggesting the component effectively captures shared behavior. In contrast, the second principal component captures roughly 10% of the variance, with its loading not reflective of a coherent set of prosocial behaviors (Appendix Figure 8). Thus, while a simplification of the inherent

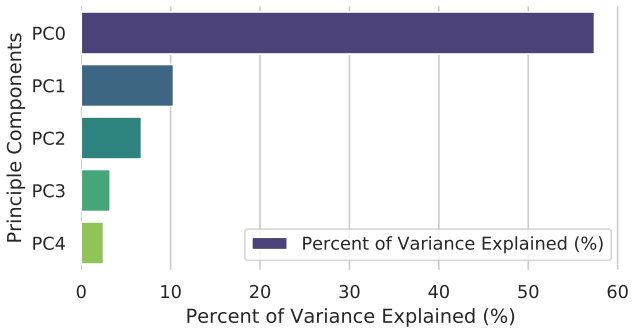

Figure 2: The percentage of variance explained by the first five principal components across the values of prosocial metrics for conversations shows that the first PCA component explains 57.4% of the variance in the prosocial metrics' values, indicating many prosocial metrics are highly correlated. The loadings for the first and second components are shown in Appendices Figures 7 and 8.

complexity of online behavior, the first principal component offers a compelling single value to act as a proxy in comparing the variety of behaviors seen in conversations. In this sense, the component acts analogously to other high-level estimates of behavior such as the toxicity score measured by Perspective API [84] that provides a single value for downstream applications.

To further test and validate the use of the first principal component as a proxy in our experiments, we calculate its MCC score with human judgments of prosociality, as was done for the individual metrics (Table 1). The resulting MCC of 0.63 is higher than the MCC for any single metric and indicates the component's value is strongly reflective of human judgments of prosociality. Thus, given the single component that explains a substantial portion of the variance and its close correlation with human judgments, we view this first component as an effective single proxy for evaluating conversations. However, we recognize that the metrics studied here, while diverse, do not capture all of the prosociality, nor does the first component capture all of the variance, so that while strongly correlated, this first component is only an initial step at estimating prosociality.

5 FORECASTING PROSOCIAL BEHAVIOR

Given that the proposed metrics reflect human judgments of prosocial behavior, we introduce computational models to forecast the degree of prosociality a conversation will ultimately from its initial discussion. Our ultimate motivation is to train a forecasting model on conversational outcomes using massive amounts of data labeled with computational-estimated prosociality and then, in §6, fine-tune this model to rank conversations based on human judgments.

5.1 Task and Experimental Setup

The first principal component strongly correlates with human judgments (§4) and therefore we treat the first component's value as a

single numeric estimate of the prosociality of a conversation following a Tlc. We refer to this value as the Tlc's *prosocial trajectory*. While using a single value to capture prosociality across all comments under a Tlc likely simplifies some nuances of different behaviors, the single value nonetheless provides a useful proxy for conversational quality akin to other antisocial metrics; further, the PCA analysis showed that a single component captured the majority of the variance, with no other component having a consistent or substantial loading on prosociality, suggesting that a single metric, while simplifying, could still be effective at reflecting broad trends in conversational prosociality.

Models are trained to predict the prosocial trajectory value given (i) the title and text of a post, (ii) the Tlc, and (iii) metadata for the comment including the subreddit and time when it was posted. Models are fit and tested using the training, development, and test partitions shown in Table 2 using MSE as the objective.

5.2 Features

Models are trained using four categories of features. The first includes features from all the prosocial metrics in §2, with the exception of accommodation; these features provide some estimate of whether the conversation is beginning on a positive note. The second category includes features reflecting the comment's relationship to the post: i) topic distributions of the post and Tlc, ii) cosine similarity of the two topic distributions, and iii) Jaccard similarity of non-stop word content in post and Tlc. For topics, a 20-topic LDA model was trained for post and Tlc text each using Mallet [54]. The third category includes features of the Tlc: i) number of words, ii) sentiment iii) subjectivity, iv) number of misspelled words, v) Flesch-Kincaid reading score, and vi) author gender. Finally, the fourth category reflects the circumstances in which the Tlc was made: i) the subreddit containing the Tlc ii) time features for the day of a month, day of a week, and hour of a day, and iii) minutes between the post's creation time and Tlc's creation time.

5.3 Models

Baselines. As a first baseline, we train a linear regression with L2 loss over all the features in §5.2, using unigram and bigram features for the Tlc text. The second baseline uses XGBoost [20], which allows us to test for combinatorial interactions between features. XGBoost was trained with a tree-based booster that had a learning rate η of 0.05, L2 regularization λ of 3.0, and L1 regularization α of 1.0. The minimum loss reduction γ required to make a further partition on a leaf node of the tree was set to 1. The maximum depth of a tree was 4. The subsample ratio of the training instances and that of columns when constructing each tree were both 0.8. We trained the model for 5000 iterations, with one parallel tree constructed during each iteration.

Our Models. We introduce two trajectory-forecasting models built on top of the Albert model [50], which is a refinement of the BERT pretrained language model [26]. In this model shown in Figure 3, the post and Tlc are fed into separate Albert-based networks and the `[CLS]` tokens from each are used as representations of their text. This vector is concatenated with a vector containing the non-textual features from §5.2 to represent the entire input. The output layer consists of a linear layer. The subreddit is represented as an embedding; these embeddings are initialized from the 300-dimensional embeddings from Kumar et al. [49] but reduced to 16 dimensions using PCA, which accounted for 72% of the variance. A total of 5278 of our 11,993 subreddits had predefined embeddings from this process, with the remaining using random initialization. To measure the effect of pre-training, we include a version of the model using only the off-the-shelf Albert parameters that are left unchanged and a second version that is first fine-tuned on Reddit post and Tlc text (respectively) using masked language modeling and then its parameters are updated during trajectory training. Additional details on hyperparameters and training are in Appendix C.

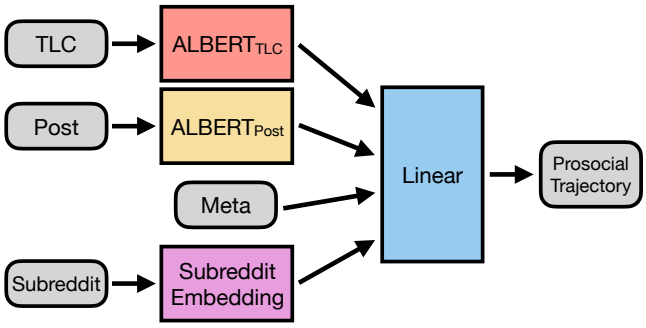

Figure 3: The proposed model for predicting a conversation's prosocial trajectory.

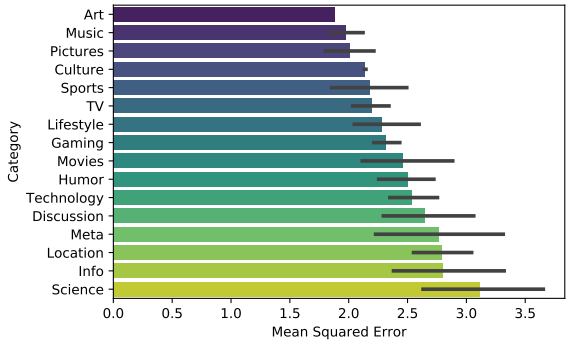

Figure 4: The MSE of prosocial forecasts within different subreddit categories shows that our Albert model attains higher performance in communities whose discussion relates to pop-culture such as Movies, Art, and Culture. Mean scores for our model and XGBoost are reported in Table 5.

5.4 Results

Models were able to identify sufficient signals of the conversational trajectory from just the post and Tlc to forecast its eventual value, as shown in Table 3. While performance is low, high performance is not expected in this setting as models only have access to the start of a conversation, which can take many potential trajectories. Nevertheless, the substantial improvement of the XGBoost and our full model over both the mean-value and linear regression baselines indicate that some signals can be reliably

found which would aid in proactive conversation sorting. Examining relative differences between the deep learning models, fine-tuning the language model parameters was critical to performance improvement, with the baseline Linear Regression outperforming the substantially more complicated model that used off-the-shelf parameters.

Conversations in some topics may be easier to forecast than others. To test for topical effects, we use the subreddit categorization from http://redditlist.com/ and compute within-category MSE. A clear trend emerges where both the XGBoost and our model performed better than average for categories related to pop-culture such as Movies, Art, and Culture ($p<0.01$ using the Kolmogorov-Smirnov test on error distributions). Figure 4 shows the mean MSE per Tlc within each category for our model, with Appendix Table 5 reporting means for both models. In contrast, both models performed worse for Science and Information subreddits which may take on different discussion patterns. These results highlight that different communities each have their own norms, which can make estimating conversational trajectory easier.

For our model, the ten subreddits with the lowest MSE included *r/aww*, *r/funny*, and *r/NatureIsFuckingLit*, three highly popular subreddits with millions of subscribers, suggesting the model performs well in lighthearted discussions. In contrast, the ten subreddits with the highest MSE included, *r/changemyview*, *r/PoliticalDiscussion*, and *r/geopolitics*, three popular subreddits that feature long, often-contentious discussions. We speculate that conversations in those communities are more unpredictable due to the inherent tension around the topics and therefore reflect a significant challenge to forecasting models.

Prosocial behavior may occur at any point in the subsequent conversation, which creates a challenge for our model that forecasts from only the initial comment. As a follow-up analysis, we test how our model's error changes relative to when the prosocial behavior occurs. We sampled 307K conversations of even length n and measure the prosociality (via the first principal component) of the first $\frac{n}{2}$ comments and last $\frac{n}{2}$ comments, ordered temporally. We then stratify the conversations relative to whether the first half was more prosocial, less prosocial, or even. Figure 5 reveals that across conversation sizes (shown up to 20 comments), our model consistently has lower error for prosocial behavior that occurs soon after the initial Tlc. Across all sizes, conversations with early prosociality have lower error (mean MSE 5.96) than those with later prosociality (mean MSE 7.81; $p<0.01$ using Wilcoxon), suggesting that models that use increasing amounts of context beyond just a Tlc [e.g, 19, 52] would perform well. ~8% of the conversations had the same estimated prosociality in each half, which suggests that future work could identify new dimensions or refine the tools of our existing measures to better discriminate between such cases.

6 RANKING CONVERSATIONS

We have shown that the prosocial trajectory of a conversation can be forecasted from the linguistic and social signals in its first comment. Here, we test whether these models can be used to rank conversations by their potential outcome. We adopt a simplified ranking setting where a model (or human) is shown a post and two of its Tlc and asked to select which Tlc is likely to lead to a more prosocial conversation.

Model	MSE	R^2
Mean Value (baseline)	3.010	-0.003
Linear Regression	2.393	0.209
XGBoost	2.209	0.269
Our model (frozen Albert)	2.209	0.157
Our model	2.230	0.262

Table 3: Mean squared Error and R^2 for forecasting prosocial conversational trajectory shows that models can estimate the trajectory value from early signals.

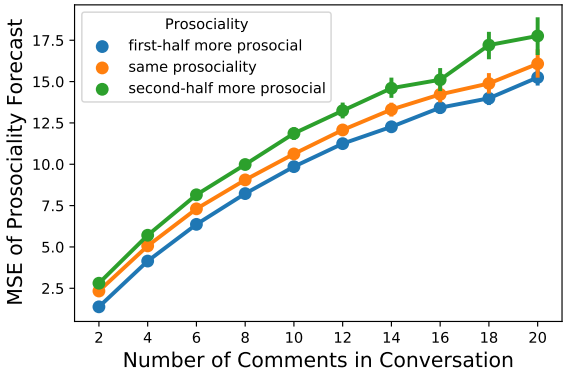

Figure 5: Error in forecasting prosociality relative to when the prosocial occurred within the subsequent conversation.

6.1 Data

Data was drawn from the 2388 instances annotated in §4 for which conversation was more prosocial.[4] An additional 1000 pairs were annotated where one of the Tlc received no replies.[5] This data was partitioned into 80% training, 10% development, and 10% test, stratifying across the three sampling strategies used to create it. Two annotators rated pairs of Tlc to the same post according to their judgments of which were more likely to result in a positive, prosocial outcome. Annotators had an Krippendorff's α=0.59. As the task is inherently difficult with no objective ground truth present in the Tlc, high agreement is not expected; however, the moderate agreement indicates that annotators were able to consistently identify a common set of lexical features they considered predictive.

6.2 Models

Our proposed ranking model, shown in Figure 6, fuses two of the post-and-comment forecasting models (Figure 3) with a linear layer to allow fine-tuning on held out human-judgments. To measure the effect of pre-training, we include a model that directly uses the trajectory estimates of the component models and selects the Tlc

[4]Note that during the previous annotation, the starts of these conversations were never shown to annotators.

[5]While empty conversations would seem to always be less prosocial, annotators preferred such conversations preferred to those containing mostly toxic comments, though such cases were rare in practice.

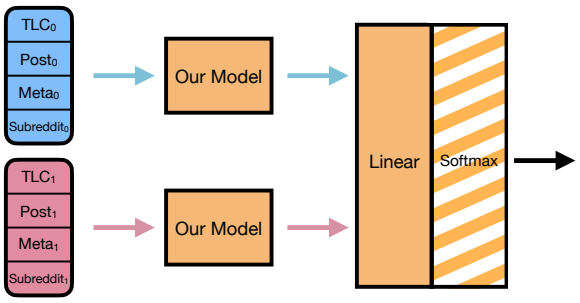

Figure 6: Diagram of our deep learning model for forecasting which conversation will be more prosocial.

with a higher estimate; a similar model is included for XGBoost. As a baseline comparison, we include a logistic regression classifier trained on unigram and bigrams directly on the TLC. Finally, we include an *oracle* classifier that uses the actual trajectory value for each TLC and selects the higher-valued; this oracle-based classifier reflects the upper bound for performance if forecasting models would perfectly estimate trajectory and rank using only that value.

6.3 Results

Models were able to surpass human performance at correctly selecting the conversation that would result in more prosocial behavior, as shown in Table 4. Consistent with prior work on predicting antisocial behavior in early conversations, high performance is not expected in this tightly-constrained setting [39, 52, 85], as new people join conversations each with their own interests and motivations that affect the trajectory. However, the moderate performance on this difficult task suggests that models can reliably pick up on prosocial signals from the very first comment in a discussion, which is sufficient for aiding in re-ranking newly-started conversations.

The pre-trained forecasting model's accuracy is the primary driver for performance. We can observe a soft upper bound for performance by comparing the model with the ranking prediction performance of a model that perfectly predicts the trajectory, shown as the *oracle trajectory prediction*. If a forecasting model would be able to forecast the trajectory with perfect accuracy (like this oracle), simply picking the conversation with the higher estimated trajectory would achieve an 86.5% accuracy at selecting the conversation that ultimately had more prosocial behavior. Although exact trajectory estimates are unlikely from the TLC alone, this illustration's result suggests that higher ranking performance is possible in future models, such as those with more data from incrementally forecasting as the conversation grows after its initial stages; indeed, models for forecasting antisocial behavior from longer context suggest that such a result is possible [19].

Despite having moderate agreement on which TLC was predicted to have a more prosocial outcome, humans performed worse than the proposed model. Annotators and the best model had only weak agreement in their judgments (α=0.29) and with 69.7% of the annotator's correct decisions also being selected by the model. This result, combined with the inter-annotator agreement, suggests that annotators were able to pick up on a complementary set of linguistic signals not used by the model, which future models might identify to improve performance.

Model	Accuracy
Logistic Regression on TLC	0.463
XGBoost	0.540
Human Prediction	0.563
Our model (only pre-trained)	0.566
Our model (fine-tuned model)	**0.578**
Oracle Trajectory Prediction	0.865

Table 4: Performances at predicting which of two conversations will have a more prosocial outcome shows that our best computation model outperforms human predictions.

7 DISCUSSION

Prosociality can take many forms and in this paper, we have developed classifiers to recognize a variety of these behaviors, showing they can be recognized and that many are correlated with each other. However, there are multiple directions that could be taken to better reflect prosociality as a whole. First, our model is agnostic to the community itself in considering what behavior should be considered prosocial, even though communities are known to have different social norms [17, 18, 53]; for example, the jocular nature of sports and gaming communities may consider politeness out of the ordinary and not inline with their desired prosocial behaviors. This direction is further supported by the observed variance in performance across different categories of subreddits (Figure 4), which suggests that directly incorporating the norms of specific communities could improve performance. Second, while our PCA score unifies many prosocial behaviors under a single metric (much like general "toxicity" scores), a significant amount of variation remains unexplained. The PCA value used in this paper offers a compelling and practical operationalization. However, further analyses are needed to identify other prototypical forms of prosociality and their effect on conversations. Third, our forecast and ranking models simplify the task to only looking at a TLC in predicting conversational trajectory. As conversations can often take unpredictable turns, these later comments are likely to influence its prosocial trajectory, which cannot be observed from the TLC alone. However, given our promising results on just the TLC, later models may improve upon these results through iteratively predicting prosocial trajectory from the growing sequence of comments in a conversation, as others have done in forecasting antisocial behavior [19, 52].

8 CONCLUSION

Online conversations can take many trajectories, not all of them pleasant. Improving our ability to recognize and highlight nascent *prosocial* conversations early on in their discussion can directly impact the daily lives and discussions of millions by fostering a more amicable and productive discourse online. This paper has introduced a series of metrics for different forms of prosocial behavior and accompanying computational techniques for recognizing them, showing that these behaviors are strongly correlated with

human judgments of prosocial conversations—and that prosociality isn't simply explained by the absence of antisocial behavior. In two experiments, we introduce a series of deep learning models showing that prosocial trajectories can be forecasted from just the initial comment in a conversation. Then, we show that these models can be adapted to predict which of two conversations is more likely to have a prosocial outcome from these signals, providing a ranking mechanism for increasing the visibility of conversations likely to have prosocial outcomes. While forecasting from such little data is difficult—but critical to ranking new conversations—our model is able to substantially improve on human performance over chance (24%) at selecting the one with better outcome. Our model provides an initial starting point for conversation ranking and we show that if the forecast was completely accurate, such models would have an upper limit of 86.5%, further motivating work in this area. Code, data, and models are available at https://github.com/davidjurgens/prosocial-conversation-forecasting.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grants No 1850221 and 2007251 and the NVIDIA Corporation through a GPU grant.

REFERENCES

[1] George A Akerlof and Rachel E Kranton. 2000. Economics and identity. *The Quarterly Journal of Economics* 115, 3 (2000), 715–753.
[2] Sara B Algoe. 2012. Find, remind, and bind: The functions of gratitude in everyday relationships. *Social and Personality Psychology Compass* 6, 6 (2012), 455–469.
[3] Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of SIGIR*. 45–54.
[4] Jo-Anne Bachorowski and Michael J Owren. 2001. Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect. *Psychological Science* 12, 3 (2001), 252–257.
[5] JinYeong Bak, Chin-Yew Lin, and Alice Oh. 2014. Self-disclosure topic model for classifying and analyzing Twitter conversations. In *Proceedings of EMNLP*. 1986–1996.
[6] Daniel Bar-Tal. 1976. *Prosocial behavior: Theory and research*. Hemisphere Publishing Corp.
[7] Azy Barak and Orit Gluck-Ofri. 2007. Degree and reciprocity of self-disclosure in online forums. *CyberPsychology & Behavior* 10, 3 (2007), 407–417.
[8] Monica Y Bartlett and David DeSteno. 2006. Gratitude and prosocial behavior: Helping when it costs you. *Psychological science* 17, 4 (2006), 319–325.
[9] C Daniel Batson and Adam A Powell. 2003. Altruism and prosocial behavior. *Handbook of psychology* (2003), 463–484.
[10] Roger Berry. 2009. You could say that: the generic second-person pronoun in modern English. *English Today* 25, 3 (2009), 29–34.
[11] Prakhar Biyani, Cornelia Caragea, Prasenjit Mitra, and John Yen. 2014. Identifying emotional and informational support in online health communities. In *Proceedings of COLING*.
[12] Jonathon D Brown and S Smart. 1991. The self and social conduct: Linking self-representations to prosocial behavior. *Journal of Personality and Social psychology* 60, 3 (1991), 368.
[13] Penelope Brown. 2015. Politeness and language. In *The International Encyclopedia of the Social and Behavioural Sciences (IESBS),(2nd ed.)*. Elsevier, 326–330.
[14] Taina Bucher. 2012. Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. *New media & society* 14, 7 (2012), 1164–1180.
[15] Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and Joao Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of EMNLP*.
[16] Anthony L Burrow and Nicolette Rainone. 2017. How many likes did I get?: Purpose moderates links between positive social media feedback and self-esteem. *Journal of Experimental Social Psychology* 69 (2017), 232–236.
[17] Stevie Chancellor, Andrea Hu, and Munmun De Choudhury. 2018. Norms matter: contrasting social support around behavior change in online weight loss communities. In *Proceedings of CHI*. 1–14.
[18] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of CSCW* (2018).
[19] Jonathan P Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In *Proceedings of EMNLP*.
[20] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of KDD*.
[21] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of CSCW*. ACM, 1217–1230.
[22] 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*. 1514–1525.
[23] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of WWW*. ACM, 699–708.
[24] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of ACL*.
[25] Sebastian Deri, Jeremie Rappaz, Luca Maria Aiello, and Daniele Quercia. 2018. Coloring in the links: Capturing social ties as they are perceived. *Proceedings of CSCW* (2018).
[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.
[27] Steve Duck. 2007. *Human relationships*. Sage.
[28] Maeve Duggan. 2017. Online harassment 2017. (2017).
[29] Robert A Emmons and Michael E McCullough. 2004. *The psychology of gratitude*. Oxford University Press.
[30] Facebook Research. 2019. Instagram Request for Proposals for Well-being and Safety Research. https://research.fb.com/programs/research-awards/proposals/instagram-request-for-proposals-for-well-being-and-safety-research/.
[31] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 85.
[32] Jeremy A Frimer, Karl Aquino, Jochen E Gebauer, Luke Lei Zhu, and Harrison Oakes. 2015. A decline in prosocial language helps explain public disapproval of the US Congress. *Proceedings of the National Academy of Sciences* 112, 21 (2015), 6591–6594.
[33] Jeremy A Frimer, Nicola K Schaefer, and Harrison Oakes. 2014. Moral actor, selfish agent. *Journal of personality and social psychology* 106, 5 (2014), 790.
[34] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
[35] Lorenz Goette, David Huffman, and Stephan Meier. 2012. The impact of social ties on group interactions: Evidence from minimal groups and randomly assigned real groups. *American Economic Journal: Microeconomics* 4, 1 (2012), 101–15.
[36] David Greatbatch and Timothy Clark. 2003. Displaying group cohesiveness: Humour and laughter in the public lectures of management gurus. *Human relations* 56, 12 (2003), 1515–1544.
[37] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. 2018. All You Need is" Love" Evading Hate Speech Detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*. 2–12.
[38] Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of ICWSM*.
[39] Yunhao Jiao, Cheng Li, Fei Wu, and Qiaozhu Mei. 2018. Find the conversation killers: A predictive study of thread-ending posts. In *Proceedings of the Web Conference*.
[40] Adam N Joinson. 2001. Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European journal of social psychology* 31, 2 (2001), 177–192.
[41] Adam N Joinson and Carina B Paine. 2007. Self-disclosure, privacy and the Internet. *The Oxford handbook of Internet psychology* 2374252 (2007).
[42] Sidney M Joward. 1971. *Self-dislosure: An Experimental Analysis of the Transparent Self*. Wiley Interscience.
[43] David Jurgens, Eshwar Chandrasekharan, and Libby Hemphill. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.
[44] Roberta L Knickerbocker. 2003. Prosocial behavior. *Center on Philanthropy at Indiana University* (2003), 1–3.
[45] Varada Kolhatkar and Maite Taboada. 2017. Constructive language in news comments. In *Proceedings of the Workshop on Abusive Language Online*.
[46] Varada Kolhatkar, Nithum Thain, Jeffrey Sorensen, Lucas Dixon, and Maite Taboada. 2020. Classifying Constructive Comments. *First Monday*. (2020).
[47] Wojciech Kulesza, Dariusz Dolinski, Avia Huisman, and Robert Majewski. 2014. The echo effect: The power of verbal mimicry to influence prosocial behavior. *Journal of Language and Social Psychology* 33, 2 (2014), 183–201.

[48] Srijan Kumar, Justin Cheng, and Jure Leskovec. 2017. Antisocial behavior on the web: Characterization and detection. In *Proceedings of the 26th International Conference on World Wide Web Companion*. 947–950.
[49] Srijan Kumar, Xikun Zhang, and Jure Leskovec. 2019. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of KD*. ACM, 1269–1278.
[50] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).
[51] David Lazer. 2015. The rise of the social algorithm. *Science* 348, 6239 (2015), 1090–1091.
[52] Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. 2018. Forecasting the presence and intensity of hostility on Instagram using linguistic and social features. In *Proceedings of ICWSM*.
[53] J Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences* 116, 20 (2019), 9785–9789.
[54] Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. (2002). http://mallet.cs.umass.edu.
[55] Michael E McCullough, Shelley D Kilpatrick, Robert A Emmons, and David B Larson. 2001. Is gratitude a moral affect? *Psychological bulletin* 127, 2 (2001), 249.
[56] Michael E McCullough, Marcia B Kimeldorf, and Adam D Cohen. 2008. An adaptation for altruism: The social causes, social effects, and social evolution of gratitude. *Current directions in psychological science* 17, 4 (2008), 281–285.
[57] David N Milne, Glen Pink, Ben Hachey, and Rafael A Calvo. 2016. Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. 118–127.
[58] Paul Mussen and Nancy Eisenberg-Berg. 1977. *Roots of caring, sharing, and helping: The development of pro-social behavior in children*. WH Freeman.
[59] Courtney Napoles, Aasish Pappu, and Joel Tetreault. 2017. Automatically identifying good conversations online (yes, they do exist!). In *Proceedings of ICWSM*.
[60] Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosato, and Brian Provenzale. 2017. Finding good conversations online: The Yahoo News annotated comments corpus. In *Proceedings of the 11th Linguistic Annotation Workshop*.
[61] Amit Navindgi, Caroline Brun, Cécile Boulard Masson, and Scott Nowson. 2016. Steps toward automatic understanding of the function of affective language in support groups. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*. 26–33.
[62] Kate G Niederhoffer and James W Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology* 21, 4 (2002), 337–360.
[63] Ariana Orvell, Ethan Kross, and Susan A Gelman. 2017. How "you" makes meaning. *Science* 355, 6331 (2017), 1299–1302.
[64] Michael J Owren and Jo-Anne Bachorowski. 2003. Reconsidering the evolution of nonlinguistic communication: The case of laughter. *Journal of Nonverbal Behavior* 27, 3 (2003), 183–200.
[65] Mitchell J Prinstein and Antonius HN Cillessen. 2003. Forms and functions of adolescent peer aggression associated with high levels of peer status. *Merrill-Palmer Quarterly (1982-)* (2003), 310–342.
[66] Sarah T Roberts. 2014. *Behind the screen: The hidden digital labor of commercial content moderation*. Ph.D. Dissertation. University of Illinois at Urbana-Champaign.
[67] Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and Psychological Effects of Hateful Speech in Online College Communities.. In *Proceedings of Web Science*.
[68] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of ACL*.
[69] Lauren E Scissors, Alastair J Gill, and Darren Gergle. 2008. Linguistic mimicry and trust in text-based CMC. In *Proceedings of CSCW*.
[70] Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *Proceedings of EMNLP*.
[71] Lee Sproull. 2011. Prosocial behavior on the net. *Daedalus* 140, 4 (2011), 140–153.
[72] Paul J Taylor and Sally Thomas. 2008. Linguistic style matching and negotiation outcome. *Negotiation and Conflict Management Research* 1, 3 (2008), 263–281.
[73] Jean M Twenge, Roy F Baumeister, C Nathan DeWall, Natalie J Ciarocco, and J Michael Bartels. 2007. Social exclusion decreases prosocial behavior. *Journal of personality and social psychology* 92, 1 (2007), 56.
[74] Twitter. 2018. Twitter health metrics proposal submission. https://blog.twitter.com/en_us/topics/company/2018/twitter-health-metrics-proposal-submission.html.
[75] Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *ACL*.
[76] Rob Voigt, Nicholas P Camp, Vinodkumar Prabhakaran, William L Hamilton, Rebecca C Hetey, Camilla M Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences* 114, 25 (2017), 6521–6526.
[77] Yafei Wang, John Yen, and David Reitter. 2015. Pragmatic alignment on social support type in health forum conversations. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*. 9–18.
[78] Zijian Wang and David Jurgens. 2018. It's going to be okay: Measuring Access to Support in Online Communities. In *Proceedings of EMNLP*.
[79] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language*.
[80] Thomas Ashby Wills. 1991. Social support and interpersonal relationships. (1991).
[81] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of EMNLP*.
[82] Michelle F Wright and Yan Li. 2011. The associations between young adults' face-to-face prosocial behaviors and their online prosocial behaviors. *Computers in Human Behavior* 27, 5 (2011), 1959–1962.
[83] Michelle F Wright and Yan Li. 2012. Prosocial behaviors in the cyber context. In *Encyclopedia of cyber behavior*. IGI Global, 328–341.
[84] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. [n.d.]. Ex machina: Personal attacks seen at scale. In *Proceedings of the Web Conference*.
[85] Justine Zhang, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of ACL*.
[86] Naitian Zhou and David Jurgens. 2020. Condolences and Empathy in Online Communities. In *Proceedings of EMNLP*.

A PROSOCIAL METRICS

This section describes the features, training, and setup for classifiers and regressors that estimate specific prosocial metrics.

A.1 Information Sharing

Information-sharing comments were identified using a classifier trained on heuristically-labeled data. Positive examples of information sharing were drawn from 18 question-focused subreddits where individuals post questions and receive potentially-informative replies (e.g., r/whatisthisthing and r/AskReddit); these subreddits cover multiple topics to prevent overfitting to sharing just one type of information. Information sharing comments were drawn from January–March of 2018 from posts that contained at least one question; replies to these questions receiving a score >2 were taken as positive examples of information sharing. Negative examples were drawn from a random sample of English-language replies to posts *not* in these subreddits. Our dataset consists of 55,542 informative comments and 300,226 comments from non-informative communities. This class skew was intentionally left imbalanced to simulate the real-life scenario where most comments are not *information-sharing*. We fit a logistic regression classifier on the unigram and bigram features with five-fold cross-validation. The hyperparameter for the minimum n-gram frequency was varied between values $[100, 50, 25, 15, 10]$. The final model obtained an F1-score of 0.713. We selected a decision threshold of ≥ 0.7 for being information sharing to reduce false positives.

We further labeled a comment as information-sharing if it contained a URLs to websites that are commonly used as information references: wikipedia.org, stackoverflow.com, quora.com, imdb.com, webmd.com, merriam-webster.com, nih.gov, weather.com, genius.

com, books.google.com, github.com, wikihow.com, answers.yahoo.com, ehow.com, thefreedictionary.com, dictionary.com, and lifehacker.com.

A.2 Laughter

Laughter is detected by identifying colloquial internet expressions signalling laughter, e.g., "haha" or "lol." As these forms may repeat or have variation, e.g., "hahhahaaha" we use a regex to detect them:

```
\ba*h+a+h+a+(h+a+)*?h*\b|\bl+o+l+(o+l+)*?\b|\bh+e+h+e+(h+e+)*?h*\b
```

A.3 Mentoring

We built a classifier with positive examples of mentoring drawn from advice-based subreddits where users post questions and the community responds with advice to those questions (appearing as TLC). Negative examples were randomly drawn from a TLC made in all other subreddits. We considered communities containing the word "Advice" in their name (e.g., r/legaladvice, r/relationship_advice, and r/mechanicadvice), excluding those with the word "Bad.". Compared to *information-sharing* subreddits, answers in *mentoring* subreddits are typically subjective in nature. We generated 500,000 negative examples using reservoir sampling. We processed and purified this dataset with the same pipeline as A.1, which resulted in a 79,430 positive examples of mentoring and 299,006 negative examples. Our logistic regression classifier for *mentoring* prediction has an F1-score of 0.762 and we manually adjust the decision threshold to $\geq$0.7 in order to reduce false positives.

A.4 Gratitude

To detect gratitude in replies, we use a fixed lexicon of words and phrases, which manual inspection showed had high precision when interpreting whether the responding user was expressing gratitude. Gratitude words are "thanks," "contented," and "blessed." Gratitude phrases are "thank you," "thankful for," "grateful for," "grateful for," "my gratitude," "i appreciate," "make me smile," "I super appreciate," "i deeply appreciate," "i really appreciate," and "bless your soul."

A.5 Esteem Enhancement

Compliments were identified using a rule-based procedure to select parts of comments referring to the user being replied to and then testing whether the sentiment around that reference was positive. An initial set of candidates was identified by looking for direct mentions of "you is/are" or "your [word] is/are." We then filter out all candidates containing where "you" is immediately preceded by "if" or "when" as analysis showed these constructions were likely to invoke the generic sense of you [10, 63] and not refer directly to the user in the parent comment. From the remaining, we extract the five words following our matched phrase and score the sentiment using VADER [38]. We use the compound sentiment score from the `vaderSentiment` library as an aggregate estimate of the positivity toward the parent comment's user. The minimum threshold for sentiment was set at 0.7 after reviewing several hundred sentiment scores showed this resulted in few false positives.

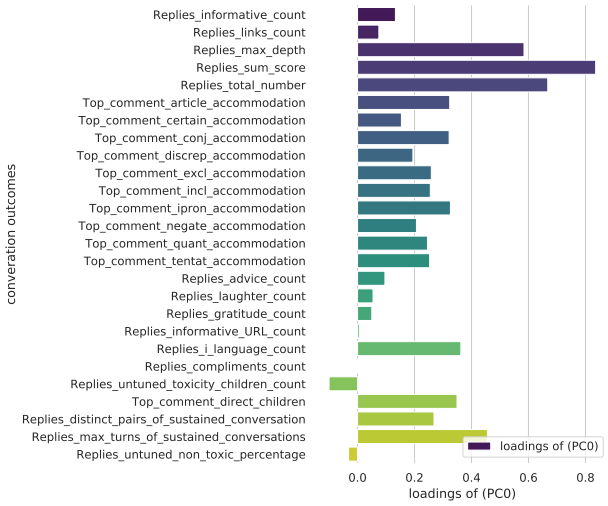

Figure 7: PCA Component 0 Loadings Across Metrics.

A.6 Donations

Fundraising and Donation behavior is measured by counting how many times a URL with one of these base domains are mentioned in the total conversation. The following URLs were drawn popular charity, fundraising, and donation organizations: gofundme.com, indiegogo.com, causes.com, kickstarter.com, patreon.com, circleup.com, lendingclub.com, fundly.com, donatekindly.org, givecampus.com, snap-raise.com, snowballfundraising.com, bonfire.com, crowdrise.com, dojiggy.com, mightycause.com, depositagift.com, wemakeit.com, donorschoose.org, fundrazr.com, rallyme.com, startsomegood.com, diabetes.org, humanesociety.org, cancer.org, nwf.org, worldwildlife.org, habitat.org, oxfam.org, unicefusa.org, wish.org, nature.org, aspca.org, savethechildren.org, wfp.org, hrc.org, hrw.org, nationalmssociety.org, redcross.org, mentalhealthamerica.net, amnesty.org, heart.org, crs.org, kiva.org, fsf.org, rotary.org, alz.org, doctorswithoutborders.org, unitedway.org, and cancer.org.

A.7 Politeness

Two prior datasets exist with politeness ratings. The data of Danescu-Niculescu-Mizil et al. [24] contains z-scored ratings of politeness for questions, whereas the data for Wang and Jurgens [78] contains ratings for statements of a variety of lengths rated on a scale in [1,5] where 3 indicates neither polite nor impolite. To build a robust classifier for Reddit, we combine both datasets, and rescale both datasets to be within [-1,1]. To obtain the politeness regressor, we first pre-train a BERT-based model [26] on Reddit data using masked language modeling. Then, we fine-tune those parameters using the Adam optimizer with a learning rate of 0.00002. The max sequence length is set to 128. While training, we adopt *MSE* as the loss function and a five-fold cross validation strategy is utilized when evaluating model's performance. Each model was run at most 5 epochs and we took the one whose average Pearson r across five folds was the highest for further usage. The final model obtained a r=0.66 with human judgments from both datasets.

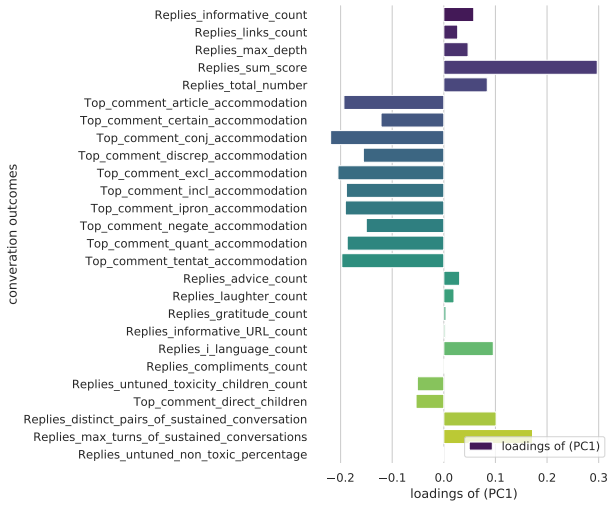

Figure 8: PCA Component 1 Loadings Across Metrics.

A.8 Supportiveness

The supportiveness regressor was trained in a similar manner as the politeness regressor, but used the only available dataset of Wang and Jurgens [78] for estimating support. Support is scored within [-1,1] with a rating of 0 indicating neither supportive nor unsupportive. A BERT model is first pre-trained on Reddit data using masked language modeling and then five-fold cross-validation is done where each fold is fine-tuned on these support ratings. We select the model with the highest rating across folds. The final model reached r=0.58 with human judgments in their data, which surpasses the state-of-the-art model results reported in Wang and Jurgens [78] for their best model.

B ADDITIONAL PCA ANALYSIS

Multiple prosocial behaviors may occur in the same conversation and to capture their regular co-occurrence, we use Principal Component Analysis (PCA) to identify the main forms of variation. PCA is computed on a matrix where each conversation is a row and the columns contain the value of each prosocial metric Shown in Figure 2 (main paper), the first principle component explains 57.4% of the variance in the data, with all other components explaining far less. The loadings of this first principle component (Figure 7) shows that this component is loading on all of the prosocial behaviors (and negatively on the antisocial behaviors) indicating that it effectively summarizes our studied prosocial behaviors within a single metric. As a comparison, we show loadings for the second largest component in Figure 8, which explains ~10% of the variance; this component does not have any clear association with prosocial behavior and seems to match conversations with high scores but little conversation. Similar trends were observed for all other components, which lacked a clear association with prosocial behavior, suggesting that a single metric can be a reasonable proxy for summarizing the prosocial behaviors.

	Our Model	XGBoost		Our Model	XGBoost
Art	1.88	1.75	Pictures	2.05	1.93
Culture	2.14	2.01	Music	2.00	2.05
TV	2.16	2.09	Lifestyle	2.16	2.19
Sports	2.27	2.19	Movies	2.50	2.29
Gaming	2.32	2.30	Humor	2.54	2.39
Technology	2.55	2.50	Meta	2.77	2.59
Discussion	2.70	2.62	Location	2.70	2.63
Info	2.85	2.79	Science	3.07	2.97

Table 5: The MSE of prosocial forecasts within different subreddit categories shows that our two top models both attain higher performance in communities whose discussion relates to pop-culture such as Movies, Art, and Culture.

Hyperparameter	Value
booster type	gbtree
learning rate η	0.05
minimum loss reduction γ	1.0
maximum depth of a tree	4
minimum child weight	1.0
subsample ratio of training instances	0.8
the subsample ratio of columns (colsample_bytree)	0.8
L2 regularization term on weights λ	3.0
L1 regularization term on weights α	1.0
number of parallel trees	1
number of boost rounds	5000
number of early stopping rounds	50

Table 6: Hyperparameters of the XGBoost Model

Hyperparameter	Our Model	Frozen Albert
learning rate	1e-5	1e-4
number of epochs	2	5
L2 penalty (c)	1e-6	1e-6
pretrained albert type (for TLC texts)	albert-base-v2	albert-base-v2
pretrained albert type (for post texts)	albert-base-v2	albert-base-v2
dropping probability of the dropout layer	0.5	0.5
subreddit embedding dimension	16	16
learning rate scheduling	linear	linear
optimizer	AdamW	AdamW
random seed	42	42

Table 7: Hyperparameters of our model (left) and the model using frozen weights for the base Albert Model (right).

C MODEL HYPERPARAMETERS

XGBoost. Hyperparameters for the XGBoost model are shown in Table 6. We tuned the learning rate η through grid search (log-linearly) in the ranges between 0.1 and 0.001. We regard models that have the lowest mean square loss as the best model. The model was trained on cpu-s for 27 hours, 21 minutes and 3 seconds, validated on the validation set every 10 iterations. The mean square error of the above model on the validation set is 1.49, and its R^2 is 0.27.

Our Albert-based Model. Hyperparameters for our Albert-based model are shown in Table 7. We tuned the learning rate and weight decay using grid search (log-linearly) in the ranges from 0.1 to 0.001, and from 1e-4 to 1e-7 respectively. We regarded models that