Social Norms on Reddit: A Demographic Analysis[*]

Sara De Candia
Polytechnic of Turin, Italy
sarad.candia@gmail.com

Gianmarco De Francisci Morales
Centai, Italy
gdfm@acm.org

Corrado Monti
Centai, Italy
me@corradomonti.com

Francesco Bonchi
Centai, Italy
Eurecat, Spain
bonchi@gmail.com

ABSTRACT

Social norms, the shared informal rules of acceptable behavior, drive and reflect the evolution of societies. As an increasingly large part of social interactions happens online, social media data offers an unprecedented opportunity to assess the perception of social norm boundaries *in-the-wild*. In this regard, Reddit's `r/AITA` represents an invaluable source of codified social norms. This subreddit is an online forum where individuals describe how they acted in a specific situation in the past, and ask for the feedback of the community about whether their behavior was deviant or socially acceptable. Other users in the community share their views and express a judgment codified by a tag.

This study focuses on assessing which factors are associated with judgements expressed by the community. Specifically, we investigate two main factors: the demographics of the author of the submission and the topic of the submission. Our analysis shows a clear gender imbalance in the judgements, with submissions by male authors receiving negative judgements with a 62% higher likelihood. Older authors (≥ 28) also have a higher chance of receiving negative judgements (+21%). Regarding topics, submissions about romantic relationships and work tend to be judged more positively (+69% and +70%, respectively), thus hinting towards a role of the community as a support group, especially for female participants. We then focus on controversial submissions which garner heterogeneous judgements. We find that these submissions are clearly separable from those ones that are unanimously judged, and that male and older (≥ 28) authors are more likely to describe controversial situations that split the community (+26% and +22%, respectively).

Finally, we focus on the characteristics of the evaluators. We find that their judgements are associated with the other communities they belong to (signifying other interests and experiences), with an effect size comparable to the demographic group of the author. By combining all these variables—demographics of the author and communities of the evaluator—we are able to build a classifier that predicts a deviance judgement on a submission with AUC = 0.85.

ACM Reference Format:
Sara De Candia, Gianmarco De Francisci Morales, Corrado Monti, and Francesco Bonchi. 2022. Social Norms on Reddit: A Demographic Analysis. In *14th ACM Web Science Conference 2022 (WebSci '22), June 26–29, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3501247.3531549

1 INTRODUCTION

Social norms are the informal rules that govern behavior in groups and societies. When social norms are internalized, abiding to them is perceived as "good" behavior, while people associate feelings of guilt or shame to behaving in a *deviant* way [25, 27]. Being informal, social norms are sometimes blurry. For this reason, the perception of their boundaries might differ across people, and individuals might be unsure about the expected behavior in specific situations.

Social norms have been extensively studied in the social sciences [9, 12, 24]: much is known about their formation, persistence, evolution, function, effects, and their link to social identity [29]. While classic studies were conducted mostly by means of questionnaires, nowadays, thanks to the fact that an increasingly large part of social interactions happens online, we can exploit social media data to assess the perception of social norm boundaries in-the-wild.

The subreddit `r/AmItheAsshole` (`r/AITA` for short) is a community on Reddit dedicated to asking for feedback about social behavior. Users describe—in a *submission*—a challenging situation they experienced in the past and how they behaved in it, and ask the community to judge their behavior. Other users in the community reply by giving their judgement, explaining their reasoning, and by leaving a codified *tag*, from a predefined set of five tags (described in Table1) which puts the blame on one of the participants in the situation: the author of the submission, or the other people involved. These judgements are an invaluable source of codified social norms.

The present study focuses on exploring which factors shape the distribution of judgements expressed by the community. In particular, we investigate two main factors: the demographics of the author of the submission and the topic of the submission. We also explore how these variables are associated with a higher level of controversy (i.e., disagreement) in the set of judgements. Finally, we switch attention to the users expressing judgements (*evaluators*), and how the communities they belong to, as a proxy for their interests, affect their feedback.

Our results show a gender imbalance in the judgments of the community: male authors receive negative (i.e., deviant) judgements with a 62% higher likelihood. Additionally, older authors also have

[*] This work was done while the authors were at ISI Foundation.

WebSci '22, June 26–29, 2022, Barcelona, Spain

ACM ISBN 978-1-4503-9191-7/22/06…$15.00
https://doi.org/10.1145/3501247.3531549

Table 1: Tags used as judgement in r/AITA, their prevalence, and their meaning as per the subreddit guidelines.

Tag	Frequency	Meaning	Explanation
NTA	57.4%	Not The A-hole	The author is NOT to blame and the other party described is to blame.
YTA	25.6%	You're The A-hole	The author is at fault in their situation.
NAH	10.2%	No A-holes Here	Neither party is to blame. All parties' actions are justified.
ESH	6.7%	Everyone Sucks Here	Both parties involved in the scenario are to blame.
NFO	0.1%	Not Enough Info	The OP never clarifies details that would determine the true judgment.

a higher chance of receiving negative judgements, 21% more for people 28 or older (compared to people 18 or younger). However, this effect is non-linear, and shows that authors in the age bracket 22-23 receive the most favorable treatment, and are 19% more likely to receive a *positive* judgement. Male and older (≥ 28) authors also receive more split judgements (+26% and +22%, respectively), which indicates a higher level of controversy associated to the situations described by them, and the involved social norms.

We also observe topical differences in the odds of negative judgements received, whereby situations related to *romantic relationships* and *work* are judged more positively (+69% and +70%, respectively), while situations related to how to behave in *society* are judged more negatively, albeit the latter difference can be explained by the gender composition of the authors within this topic.

Finally, the judgement expressed by evaluators is associated with the communities they belong to, with a strength of association comparable to the demographics of the author. By combining the demographic information of the author and the communities of the evaluators, a machine learning model is able to predict single deviance judgements with an AUC ROC of 0.85.

Overall, our results hint at a pattern of usage of the r/AITA community as a support group for a portion of the users [6]. For example, this pattern can explain the positive tendency in judgements towards submissions dealing with romantic relationships: users seek and find encouragement rather than judgement. A similar pattern might be present in work-related submissions; this interpretation is also supported by a worker community appearing as a strong predictor of positive judgement. Indeed, several of our predictors for positive judgements by evaluators represent support communities, where people vent their frustration.

This work represents a first but important step in addressing questions regarding social norms, their perception, and possible determinants, by looking at online and social media data. Our results, while being purely observational, provide a picture of the reality of the r/AITA community, and should inform future work on teasing out possible causal effects.

2 BACKGROUND AND RELATED WORK

Social norms are shared, informal rules that define which behavior is deemed acceptable in a certain group or society [19]. As such, they have been extensively studied in different disciplines, including sociology, anthropology, psychology, and economics [17]. While some scholar interpret social norms as an individual construct, others view it as a collective construct instead, reproduced by institutions and other social groups. In this context, it is important to identify the norm's *reference group*, i.e., the individuals whose behavior and approval or disapproval define and sustain the norm [19]. In our context, therefore, we focus on Reddit users as our reference group; we refer to Duggan and Smith [11] for an in-depth analysis of the general demographics of Reddit users.

The importance of social norms lie in their power to shape behavior. One key pathway for a social norm to influence decisions and actions is by the creation of external obligations, e.g., through role modeling, social pressure, or anticipation of rewards and penalties [19]. Social norms might simply provide information about what are the reference groups' collective expectations [16]. Therefore, different societies, countries, and demographic groups may exhibit different social norms, and give different importance to their application [15, 23]. For instance, Gelfand et al. [15] compared how acceptable is deviant behavior in 33 different countries.

One key demographic trait we investigate is self-reported gender of r/AITA participants. Gender, as a social construct, is often shaped by social norms, called *gender norms*. Such norms [8] define which behavior is expected from different genders, and which actions are considered appropriate for women and men in that group or society. Of course, they vary over time, places, and cultures [22]. As such, many works in the literature have studied how social norms are perceived by different demographic groups. For instance, Pampel [23] studied how gender-egalitarian social norms are perceived by different cohorts in terms of age, location, and education levels. Doey et al. [10] studied how social norms differ between genders regarding the acceptability of shyness in children in United States. Many of these studies point to the tension between egalitarian social norms and more gender-discriminating norms, often to better understand and encourage the change of detrimental social norms.

Similarly to gender, the role of age in establishing social norms has long been subject of study [21]. For example, Settersten Jr and Hägestad [28] showed evidence of flexible deadlines for 'age appropriate' life transitions, such as leaving home, marrying, and childbearing. Chudacoff [7] revealed that age consciousness is a relatively recent development (late 19th century), when age took a central role as a measure of charting the life course.

A line of research has studied social norms and moral judgements in online communities as well. For instance, Yee et al. [32] studied behavioral social norms in the community of Second Life. The idea that expression of social norms is linked to self-validation and approval-seeking has been confirmed also for social media [4], where individuals might disclose certain real-life events in order to validate their decision-making.

In fact, the vast data publicly disclosed on the Web has reinforced the field of descriptive ethics, which aims at describing people's moral judgments—instead of focusing on theoretical prescriptions on morality. Scholars developed large datasets such as Scruples [14] to study ethical judgements over real-life anecdotes. Through this data set, it is possible to show how many situations are naturally divisive (different evaluators often give different judgements).

Here, we focus on the Reddit community r/AITA, where individuals ask and receive moral judgments on their behavior. A few previous works analyzed this community, mostly focusing on the task of predicting judgements by using natural language processing tools applied on the text of the submissions or the comments. For instance, Botzer et al. [5] trained a classifier to distinguish between comments associated with a positive judgement and those associated with a negative one, with the goal of automatically inferring the moral valence of text. Similar prediction tasks were studied by Sarat et al. [26] and Zhou et al. [33]. Specifically, they developed a machine learning model to predict which judgement will be given to a certain submission by using the text and some Reddit metadata of the author, mostly related to their upvotes and downvotes on Reddit. They find that metadata are more predictive than linguistic traits. For instance, "karma" (the total number of upvotes received by a user) and upvote-to-downvote ratio are the most predictive features. Their best-performing model is a Random Forest classifier, able to achieve an F1-score of 0.76, qualitatively similar to our results using demographic attributes and community participation. Finally, Cannon et al. [6] has studied the behavior of users of r/AITA in a holistic fashion, but without focusing on the positive and negative judgements.

Our main focus is instead on the relationship between demographic characteristics and the social norms of the subreddit, as expressed by users in their judgements. As shown by Flesch [13] it is in fact possible to collect gender information from anonymous posters through their self-declaration. Doing so can highlight gender differences in participation rates across different subreddits, as found by Thelwall and Stuart [31]. In this work, we apply a similar methodology on this data to investigate how social norms manifest on r/AITA. To the best of our knowledge, our work is the first to analyze the relationship between demographic attributes and received judgements, and shows significant differences between demographic segments in the distribution of judgements and in their level of controversy.

3 DATA

Reddit is a social news and discussion website, which has consistently ranked among the top ten most visited websites in the United States over the past years.[1] Discussions on Reddit are organized in topical communities called *subreddits*; r/AmItheAsshole (r/AITA for short) is one of them. In this community, users posts *submissions* that consist of a title and a textual content. In a submission, the *author* explains a challenging situation they were involved in, and asks the community to judge their behavior in that situation. Typically, the author is unsure about the appropriateness of their behavior in the specific social context, i.e., they feel guilt about possibly breaking a social norm with a *deviant* behavior. Often, authors also provide their age and gender (via a codified tag within the title or text, e.g., '24F') as contextual information. Other members of the community reply with comments under the submission, expand the discussion, and provide their judgement on the behavior of the author. Henceforth, we refer to users that provide a judgement as *evaluators*. Judgements are codified according to one of five tags which appear in the comment: YTA, NTA, ESH, NAH, and NFO, explained in Table 1. In particular, we recognize YTA and ESH as *negative* judgements on the behavior of the author, i.e., the evaluator judges the behavior of the author as deviant according to their perceived social norms. Conversely, NTA and NAH are *positive* judgements, i.e., the behavior of author is judged as *conforming* to social norms. Positive judgements account for 67.6% of the comments, while negative ones for 32.8%.

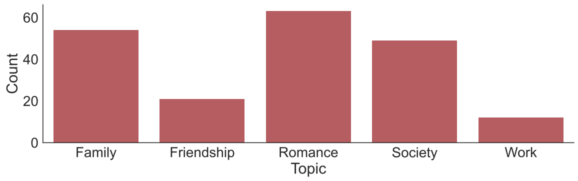


Figure 1: Distribution of hand-labelled topics in a random sample of 200 submissions.

We collect our main data set from the Pushshift Reddit data collection [3]. To do so, we gather all the submissions and comments on r/AITA in a timespan of 6 years (from the beginning of 2014 to the end of 2019), thus obtaining 354k submissions and 13M comments. After removing bot accounts,[2] the total number of evaluators in the subreddit is 499 366; however, 43% of those wrote only a single comment. Since we are interested in the behavior of the community as a whole, we focus on its most active members; as such, we restrict our attention to evaluators that have written at least 15 comments (i.e., 10% of all evaluators). The final number of evaluators in our data set is therefore 51 049, accounting for more than 4M judgements. Finally, we do not consider in our analysis the tag NFO ("not enough information"), which accounts for only 0.1% of judgments, as it does not express an actual assessment.

The final dataset contains a large number of submissions from different authors. While some of these accounts might be throwaway (i.e., they are created for the purpose of submitting to r/AITA and used only once to protect the anonymity of the author), here we are only interested in their demographic attributes rather than any information from their online profile. In fact, we focus on the self-disclosed demographic information contained within the submission, and retain for our analysis only the authors that provide such contextual information, as explained next.

3.1 Demographic information

We use a carefully-crafted regular expression to identify the self-reported age and gender of authors. For simplicity and consistency, we limit ourselves to the cases in which gender and age are both expressed at the same time. Furthermore, we consider only "M" (male) and "F" (female) as gender tags, since this is the only codified information available. Section A details the regular expression developed for this purpose.

Particular care needs to be taken to extract the correct information. For instance, authors typically report the demographic

[1] https://en.wikipedia.org/wiki/Reddit

[2] We use a combination of lists of known bots available on Reddit, pattern matching in the username, and manual cleaning.

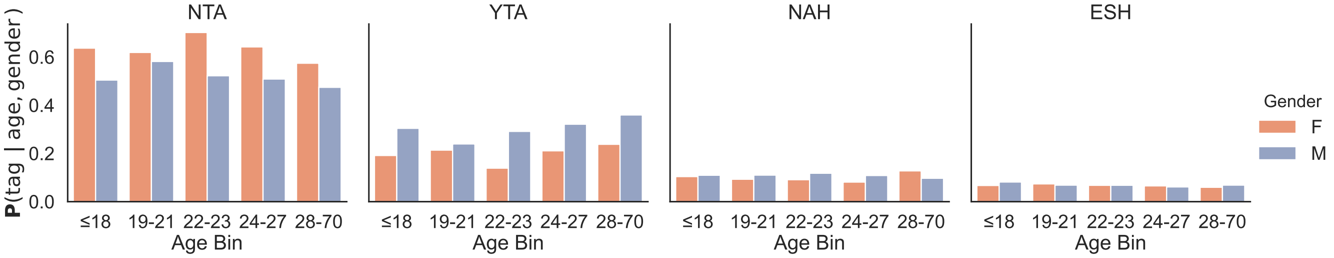


Figure 2: Conditional probability of receiving a specific judgement given the demographic group of the author.

information of all the actors in the situation described (e.g., "I (24F) live at home with my mom (56) and my brother (34)"). Fortunately, it is convention of the community to report the information of the author after a pronoun (e.g., 'I' or 'my'), which makes automatic extraction possible.

To evaluate the regular expression, we take a random sample of 100 submissions and manually check the result. We find no false positives, and only a single false negative case. In addition, when demographic information is extracted by the regular expression, it matches the manually labelled one in 100% of the cases.

We apply this regular expression to our data set, and obtain 14 126 submissions (8% of the total) with demographic information about the author. From this set, we further remove 3995 submissions that do not receive any judgement. Our final data set contains 10 131 submissions, for which age and gender information about the author is available, and 307 629 judgements, divided in NTA, YTA, NAH, and ESH. The average number of judgements per submission is 30; the median is 10. Regarding gender distribution, we find that 54% of the authors in the data set self-report as female and 46% as male, in sharp contrast with Reddit's general demographic.[3]

In order to simplify the analysis of age, we discretize this information into age groups. We divide the distribution of age of the authors in five quintiles, i.e., five approximately equally-populated bins. Each age group therefore covers a similar number of submissions. This procedure results in the following age groups: ≤18, 19-21, 22-23, 24-27, and 28-70. The unequal size (in years) of the age groups is to be expected given the younger demographic of the Reddit user base.

3.2 Topics

To further characterize the discussions within the community in terms of their topic, we inspect a random sample of 200 submissions. We use an open coding procedure from grounded theory to define the topics bottom-up. By using this procedure we identify the following five topics which cover all the submissions in the sample:

Family: situations related to relatives.
Friendship: situations related to friends.
Work: situations related to work environments.
Romance: situations related to significant others.
Society: situations concerning broad arguments such as politics, racism, and gender issues.

We associate each of the 200 submissions to its main topic among these five. Figure 1 shows the distribution of these topics in the hand-labelled submissions. Table 2 shows the per-topic distribution of self-reported demographic information in this sample of submissions.

Table 2: Self-reported age (as mean and standard deviation) and gender (as percentage of female authors) for each hand-labelled topic in our sample.

Topic	Average age	Female authors
Family	22.1 ± 5.8	68.8%
Friendship	22.0 ± 7.3	52.4%
Romance	24.7 ± 5.1	70.2%
Society	22.6 ± 5.0	41.3%
Work	22.3 ± 4.1	30.0%

4 ANALYSIS

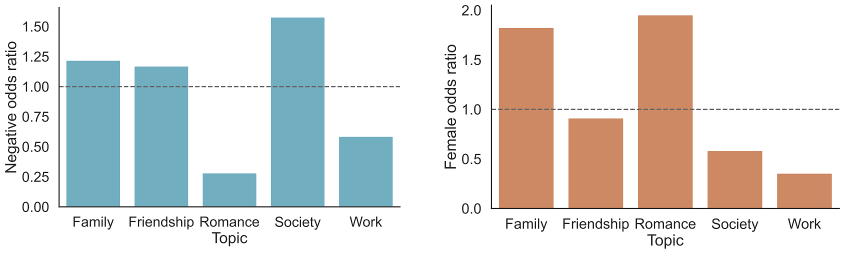
Our analysis focuses on how demographic and topical factors associate with the distribution of the received judgements, with the homogeneity of such judgements, and with the communities of the evaluators. In the rest of this section, we answer the following research questions:

- **RQ1**: What is the relationship between demographic attributes and judgements received?
- **RQ2**: Is there a topical effect on the judgements?
- **RQ3**: Is a higher level of controversy in a set of judgements associated with the demographic of the author?
- **RQ4**: Are there significant associations between the communities of the evaluator and their judgements?

4.1 Demographic attributes

Our first analysis focuses on the relationship between age of the author and the judgements they receive. We perform a χ^2 test between the categorical variable representing age bins and the tag of the judgements. The resulting p-value of this test is 0.009 and consequently it is significant at the 0.01 level. Thus, the distribution of judgements one receives differs according to their age group. Then, we focus on the association between gender of the author and the received judgements. A χ^2 test tells us that there is a significant ($p < 10^{-5}$) association between the two.

[3] https://www.statista.com/statistics/1255182



(a) Odds ratios of negative judgments for each topic compared to the global odds.

(b) Odds ratios of female authors for each topic compared to the global odds.

Topic	Family	Romance	Society	Work
Coeff.	-0.08	-1.17***	-0.00	-1.19***

(c) Coefficients of a binomial regression model to predict the number of negative judgements in a submission. We mark the level of significance with * ($p < 0.05$), ** ($p < 0.01$), or *** ($p < 0.001$).

Figure 3: Negative judgements and topics.

Figure 2 shows a global picture of how judgements are distributed across these demographic groups. From this figure, we can appreciate the direction of the differences: submissions by male authors receive a negative judgement with a higher probability than female ones; the effect on age is harder to discern, as it is not monotonic. For this reason, we employ a regression model to assess this effect.

For each tag we fit a binomial regression model which predicts the number of received judgments with the given tag. The model uses the demographic group of the author as independent variables. We code such information by using two categorical variables for gender and age group; the reference category is female and ≤ 18 of age. Table 3 presents the coefficients of the models. In accord with Figure 2, the strongest effect is due to the gender of the author: by looking at the odds ratios, male users are 73% more likely to receive a YTA judgement.[4] This model allows us to discern the direction of the bias due to age: older authors are more likely to receive a YTA judgement (10% more likely for 24-27, and 31% for 28-70). This effect is however non-linear with age: the age group with the most favorable judgements is 22-23, more than both younger and older users. Almost all the coefficients for YTA and NTA tags are significant at $p < 0.001$, while the scarcity of data for the remaining two tags (ESH and NAH) hides some of the significance. For this reason, we aggregate judgements negative towards the behavior of the author (YTA and ESH) in one category. We report the coefficients[5] for this category as last column in Table 3. The results are qualitatively similar to those of YTA.

Table 3: Coefficients (log odds ratios) of binomial regression models (one per tag) to model the number of received judgements with the given tag by using the demographic groups of the author. We include an additional model for combined negative judgements (YTA and ESH). We mark the level of significance with * ($p < 0.05$), ** ($p < 0.01$), or *** ($p < 0.001$).

Demog.	YTA	NTA	ESH	NAH	Negative
M	0.55***	−0.47***	0.03*	0.09***	0.48***
19-21	−0.10***	0.11***	−0.04	−0.06**	−0.10***
22-23	−0.22***	0.20***	−0.09***	−0.04	−0.21***
24-27	0.10***	0.02	−0.16***	−0.14***	0.04**
28-70	0.27***	−0.19***	−0.14***	0.02	0.19***

4.2 Topics

We then use the set of 200 submissions with manually labelled topic to study the association between topic and judgements. A χ^2 test tells us that there is a significant ($p < 10^{-5}$) difference in the distribution of tags among the topics. To assess this effect we look at the odds ratios of the presence of negative judgements within each topic, comparing it to the global odds. Figure 3a shows a larger presence of negative judgements for *Society* and a smaller one for *Romance* and *Work*. To see if this effect can be explained only by considering the difference in gender distribution of authors, first we plot the odds ratios for female authors in each topic in Figure 3b. Indeed, we observe a higher ratio of female authors in *Romance* which may explain the more positive judgements (as seen in the previous section). Similarly but in the opposite direction, male authors are overrepresented in *Society*, and therefore the higher number of negative judgements can be indeed explained by the results from the previous section. Conversely, the differences in *Work* and *Family* cannot be explained by gender imbalance alone.

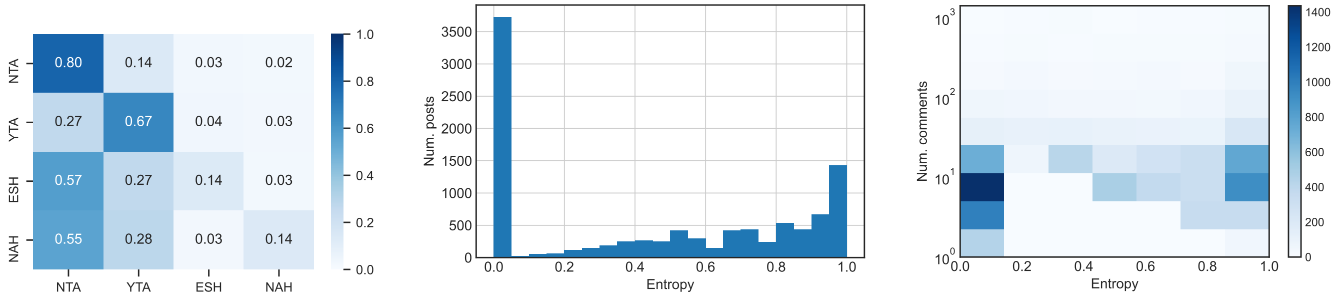
To estimate these effects, we use a binomial regression model, similarly to the previous section. The model predicts the number of negative judgements and it only uses topic and gender of author of each submission as predictors (results are qualitatively similar when including also age groups). For topics we use *Friendship* as the reference case since it shows both a gender odds and a negative judgement odds similar to the global ones. Figure 3c shows a significant effect of the topic for *Romance* and *Work*, while the effects for the other topics are explained away by gender imbalance. For *Romance* and *Work*, judgements are more positive than their gender balance would predict.

4.3 Divisiveness

An important feature of social norms is that they are informal and perceived subjectively. Consequently, some situations are bound to be divisive and controversial, and moral judgements regarding them split the community. Here we ask whether there are specific

[4] The coefficient represents the log-odds-ratio, so the odds ratio is $e^{0.55} = 1.73$.

[5] Since this operation transforms the problem into a binary classification task, positive judgements represent the reference case, and therefore their coefficients would be simply the opposite of negative ones.



(a) Each row of the matrix reports the conditional probability of finding another tag (column) given the presence of the specific tag (row) in the same submission.

(b) Distribution of the binary entropy of the judgements for a given submission (computed by aggregating tags into positive and negative). Most of the submissions have unanimous judgements, but a significant fraction is more controversial.

(c) Joint distribution of binary entropy of the judgements and number of comments of a submission. Submissions with high entropy receive a large number of comments, and, on average, more than submissions with unanimous judgements.

Figure 4: Co-occurrence of different judgements under the same submission.

demographic associations with the likelihood of a submission being controversial.

We begin our investigation by looking at the co-occurrences of tags within a submission. Figure 4a shows a significant probability mass in the off-diagonal cells. In particular, a submission with a YTA tag has a 27% probability of also receiving an NTA, while 14% of the submissions with an NTA also receive a YTA. The ESH and NAH tag present a similar conditional distribution, which does not differ significantly from their global one. For this reason, and given that ESH and NAH represent a tiny fraction of the judgements, we again aggregate them with the respective positive and negative tags.

We use this reduced dataset with a binary tag (positive or negative) to study the divisiveness of each submission. In order to quantify how controversial a submission is, we measure the *entropy* of its judgments, since it best represents the uncertainty among possible outcomes (i.e., judgements). Values closer to 1 indicate maximum uncertainty and therefore the maximum divisiveness: judgements are equally split between positives and negatives. Conversely, values closer to 0 represent the maximum level of certainty: all judgements are unanimous (either positive or negative).

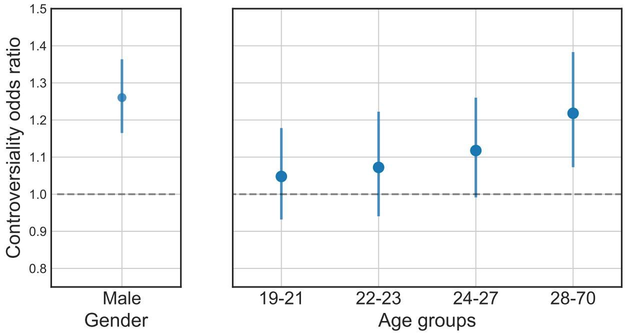


Figure 5: Regression coefficients (with 95% confidence intervals) for a logistic regression model for the divisiveness of a submission given the demographic information of its author. The label is based on the binarized entropy of the judgements received by the submission (median as threshold). Male and older authors submit more controversial situations which receive split judgements.

Figure 4b shows the distribution of this quantity in our dataset. More than a third of the submissions receive unanimous judgements, but a significant quantity produces some form of disagreement in the community. Interestingly, the number of submissions increases with the entropy, and extremely divisive ones are relatively common. One might suspect that submissions with high entropy are due to a low number of comments (e.g., a submission with a single NTA and a single YTA would have entropy 1.0), however we find this is not the case. Indeed, Figure 4c shows the joint distribution of the number of comments a submission receives, and the entropy of its judgements. Clearly controversial submissions receive a large number of judgements (35 for controversial submissions on average vs 25 for unanimous ones). By looking at the distribution, one might infer that the submissions come from two superimposed generating distributions, one for controversial ones and one for unanimous ones. We leverage this intuition to define a classification problem, so to get insights into this key feature of the community process.

We define a binary label for each submission which indicates whether its entropy is above the median of the distribution (0.49); if so, we label the submission as controversial. Therefore, by construction, we split the dataset evenly between the two classes.[6] Figure 5 shows the coefficients of a logistic regression model trained to predict this label by using the demographic information of the author of the submission. Again, we see a significant and strong imbalance with respect to gender, whereby submissions by male authors are more divisive (26% more likely to be controversial). A similar effect, albeit with a smaller effect size, can be seen for age: submissions by older authors also receive more split judgements (+22% for 28-70).

[6] Different formulations of the problem (e.g., simple and quantile regression) give qualitatively similar results.

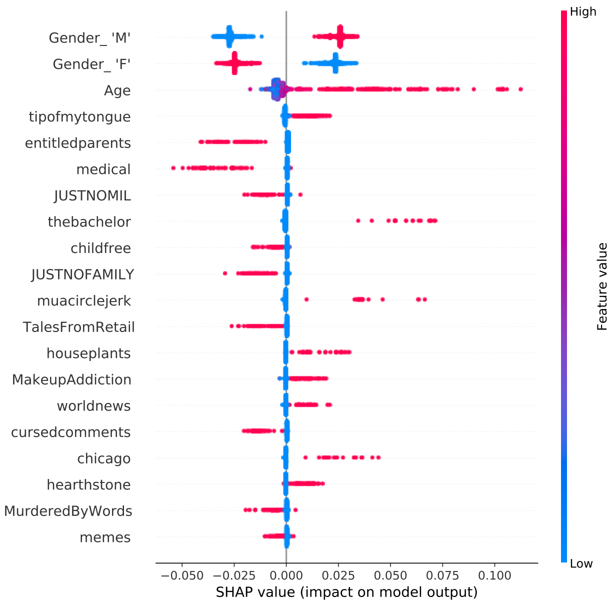


Figure 6: SHAP values plot of the most important feature in the prediction of negative judgements using demographic information of the author and communities participated by the evaluator.

Although this difference is significant only for the oldest age group, we can see a trend in the coefficients in the figure. This lack of significance is thus probably due to the small effect sizes for younger age groups. Similarly, we do not find any significant effect for the topics, due to the small size of the hand-labelled dataset.

4.4 Prediction

Finally, we ask ourselves how the interests of an evaluator (as proxied by the communities the participate in) are associated with the judgements they express. To answer this question, we build a machine learning model that, given pair of author and evaluator, predicts whether the latter will give a negative judgement to the former. For this prediction task, we use as features the age group and gender of the author, and the communities participated by the evaluator. As a proxy for participation, we use all the subreddits in which an evaluator posted at least one comment. After gathering in this way all the relevant data, we restrict our analysis to the 1000 most frequent subreddits in our data set.

In order to limit our features to the ones that are statistically relevant and to avoid leveraging spurious correlation in our data set, we perform an initial step of feature selection. In particular, we employ the family-wise error rate as a selection criterion, and keep only features that obtain a p-value lower than 0.05. As a prediction algorithm, we chose Random Forest, because of its excellent accuracy on similar tasks, the ability to model nonlinear behavior, and its ease of explainability through SHAP values [20]. In particular, preliminary experiments showed that linear models (e.g., logistic regression) perform much worse on this use case.

We evaluate results in 10-fold cross-validation: for each fold, we first apply the feature-selection algorithm to the training set; then, we use nested cross-validation to choose the hyper-parameters of the Random Forest (i.e., maximum depth and minimum number of samples for a split); finally, we test the chosen model on the remaining fold. The average number of selected features is 62. The average AUC ROC is 0.85 (Table 4).

In order to answer our research question, we resort to SHAP values. To obtain the best possible results, here we train our model on the whole data set. In this case, the family-wise error rate criterion selects 113 subreddits as significant (plus the demographic features of the author). We fix the maximum depth of the model to 8.

Figure 6 shows the SHAP values for the features identified as most important. Demographic features of the author are recognized as the most important factors. However, a number of communities participated by the evaluator are also recognized as important. For a number of communities (e.g., `entitledparents`, `medical`, `JUSTNOMIL`, `childfree`, `JUSTNOFAMILY`, `TalesFromRetail`) participation is associated to expressing more positive judgements; for others (`tipofmytongue`, `thebachelor`, `muacirclejerk`, `houseplants`, `MakeupAddiction`), it is associated to more negative judgements. We discuss these results and give a potential interpretation in the next section.

5 DISCUSSION

In this study we have explored the relationship between demographic characteristics and the perception of social norms on Reddit's `r/AITA`. When focusing on the demographic group of the author, we found an imbalance in received judgements: male and older users receive more negative ones. In addition, there are well-separated clusters of submissions in terms of divisiveness, with a large fraction of unanimous ones, but also a significant amount of controversial ones that attract a large number of evaluators. The authors of the most controversial submissions belong to the same demographic groups: older and male.

The topic of the submission is also associated with different judgement tendencies: situations dealing with societal issues receive more negative judgements, while those dealing with romance and work more positive ones. These topics are populated by different demographic groups within the community, whereby societal issues are more popular among male authors, which may explain the difference in judgements. Conversely, female authors are more represented in submissions about romantic relationships, and these submissions attract more positive judgements, even more than would be expected by their gender ratio. Interestingly, work-related submissions are created with higher probability by male authors, but at the same time, they receive more positive judgements than expected by their gender proportions.

Table 4: Results (mean and standard deviation across 10-fold cross-validation) for the prediction of negative judgements by using demographic information of the author and communities participated in by the evaluator.

Balanced Accuracy	AUC ROC
0.76 ± 0.003	0.85 ± 0.004

Finally, we analyzed the association between judgements and the communities of the evaluators that produce them. We find an effect given by these communities, whose strength is comparable to the demographic groups of the author. By combining all this information, we obtain a prediction model for a judgement given a submission with an AUC ROC of 0.85.

It is well known that people perceiving themselves deviating from social norms are subject to feelings of guilt [25, 27]. These feelings may be one of the main drivers for people to post on r/AITA, either to receive social feedback or as a coping mechanism. Therefore, one interpretation of our findings is that r/AITA and similar communities act as support group for some of their members. For example, an explanation for female individuals in our study receiving on average more positive feedback, is their usage of the community as a coping mechanism, in line with previous research about different coping strategies in genders [30]. Instead, male users might use it more as a discussion forum, with the possible exception of work-related discussions.

The usage of the community as a support mechanism could also explain, in fact, the significantly larger rate of positive judgements for submissions related to work issues. This result is in line with some of the predictors we find for the judgements made by evaluators, i.e., the other communities they participate in. For example, evaluators that participate in TalesFromRetail, which is a community about daily experiences of retail workers, tend to be more positive. More generally, several of our predictors for positive judgements can be interpreted as support communities where people vent their frustration: either explicitly as JUSTNOMiL and JUSTNOFAMILY (self-described as support communities for those with, respectively, abusive mother-in-laws and challenging family dynamics), or implicitly such as entitledparents and childfree (often used to deal with, respectively, self-centered relatives and social pressures related to reproduction). Conversely, the predictors for negative judgements include participation to thebachelor (dedicated to a dating reality show), muacirclejerk and MakeupAddiction (discussing cosmetics), and houseplants, which may be explained by demographic biases in these communities.

The different divisiveness generated by submissions by authors of different gender can have several explanations. On the one hand, psychological literature claims that male and female individuals might differ in aggressiveness [2, 18]. In our case, this claim might explain the higher level of controversy on submissions by male authors. If this were the case, we would expect a homophilic effect of gender on positive judgements, which, unfortunately, we cannot measure given the lack of data for the evaluators. On the other hand, people of different gender might be subject to different social pressures, and thus feel differently comfortable in (or compelled to) describing difficult situations in public [1]. This phenomenon might create a reporting bias, whereby male authors feel comfortable in sharing more controversial situations, which could explain our results.

As in any study dealing with social media data, there are some limitations. First, ours is an observational study, thus there might be hidden confounders that we do not take into account. For instance, there may be representation and self-selection biases that are hard to control for. Therefore, the coefficients we find *cannot be interpreted as a direct causal effect*. This study design has the advantage of a higher ecologic validity, and an easier access to large amounts of data, but presents important causal inference challenges. A way to circumvent this limitation would be to find pairs of submissions where situation described is the same or similar, and judgements differ. This quasi-experimental study design however is more involved, as it entails processing the text with advanced NLP tools that are able to assess the semantic similarity of the situations described in the submissions. This direction is a promising one for future work.

More in general, our analysis prescinds from the textual content of the submissions and of the judgements (apart form the extraction of the demographic information). Textual information is however clearly an important source of information, which we wish to exploit both for qualitative and quantitative analysis. A systematic qualitative analysis would help us understand better the kind of situations discussed in the subreddit, and thus map the boundaries of social norms in the community. While we have started this process by using open coding to map the topics of a small number of submissions, this task could benefit from further quantitative analysis and automation. For instance, this data could be used to train a topic classification model to extend the current analysis to a larger data set. However, some of the hypotheses advanced in this section still demand human validation by studying the text of the discussion, which requires significant effort.

Finally, we have provided a limited characterization of the determinants behind the judgements of the evaluators. To have a clearer picture, it would be useful to create a socio-demographic characterization of the communities (i.e., subreddits) they participate in. This problem could be tackled as a semi-supervised learning starting from self-disclosures of the evaluators.

REFERENCES

[1] Maya Asher, Anu Asnaani, and Idan M Aderka. 2017. Gender differences in social anxiety disorder: A review. *Clinical psychology review* 56 (2017), 1–12.
[2] Alvaro Q Barriga, Elizabeth M Morrison, Albert K Liau, and John C Gibbs. 2001. Moral cognition: Explaining the gender difference in antisocial behavior. *Merrill-Palmer Quarterly (1982-)* (2001), 532–562.
[3] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 830–839.
[4] Natalya N Bazarova and Yoon Hyung Choi. 2014. Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network sites. *Journal of Communication* 64, 4 (2014), 635–657.
[5] Nicholas Botzer, Shawn Gu, and Tim Weninger. 2021. Analysis of moral judgement on Reddit. *arXiv preprint arXiv:2101.07664* (2021).
[6] Emily Cannon, Bianca Crouse, Souvick Ghosh, Nicholas Rihn, and Kristen Chua. 2021. "Don't Downvote A$$$$$$$s!!": An Exploration of Reddit's Advice Communities. *arXiv preprint arXiv:2109.09044* (2021).
[7] Howard P Chudacoff. 1989. *How old are you? Age Consciousness in American Culture.* Princeton University Press.
[8] Ben Cislaghi, Karima Manji, and Lori Heise. 2018. Social norms and gender-related harmful practices: what assistance from the theory to the practice? (2018).
[9] James S Coleman. 1994. *Foundations of Social Theory.* Harvard University Press.
[10] Laura Doey, Robert J Coplan, and Mila Kingsbury. 2014. Bashful boys and coy girls: A review of gender differences in childhood shyness. *Sex Roles* 70, 7 (2014), 255–266.
[11] Maeve Duggan and Aaron Smith. 2013. 6% of online adults are reddit users. *Pew Internet & American Life Project* 3 (2013), 1–10.
[12] Émile Durkheim. 1895. *Les règles de la méthode sociologique.* Flammarion.
[13] Marie Flesch. 2019. Mapping the itineraries and interests of internet users with the RedditGender corpus. *Social Media Corpora for the Humanities* (2019), 11.
[14] Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. *arXiv preprint arXiv:2011.00620* (2020).

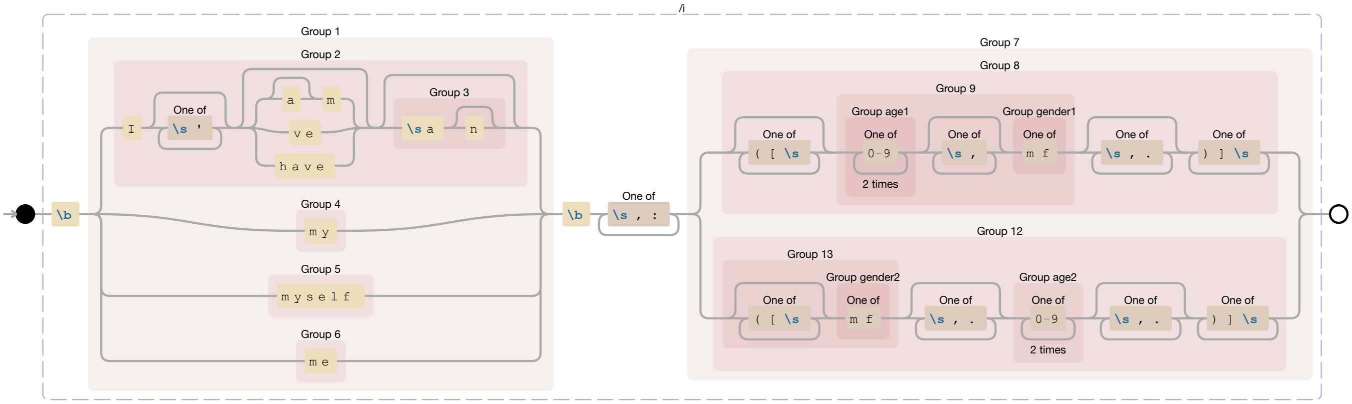


Figure 7: State diagram of the regular expression that extracts age and gender information.

[15] Michele J Gelfand, Jana L Raver, Lisa Nishii, Lisa M Leslie, Janetta Lun, Beng Chong Lim, Lili Duan, Assaf Almaliach, Soon Ang, Jakobina Arnadottir, et al. 2011. Differences between tight and loose cultures: A 33-nation study. *science* 332, 6033 (2011), 1100–1104.

[16] Jack P Gibbs. 1965. Norms: The problem of definition and classification. *Amer. J. Sociology* 70, 5 (1965), 586–594.

[17] Michael Hechter and Karl-Dieter Opp. 2001. Social norms. (2001).

[18] Janet Shibley Hyde. 2005. The gender similarities hypothesis. *American psychologist* 60, 6 (2005), 581.

[19] Sophie Legros and Beniamino Cislaghi. 2020. Mapping the social-norms literature: An overview of reviews. *Perspectives on Psychological Science* 15, 1 (2020), 62–80.

[20] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[21] Bernice L Neugarten, Joan W Moore, and John C Lowe. 1965. Age norms, age constraints, and adult socialization. *American journal of Sociology* 70, 6 (1965), 710–717.

[22] World Health Organization. 2022. Gender and health. https://www.who.int/health-topics/gender

[23] Fred Pampel. 2011. Cohort changes in the socio-demographic determinants of gender egalitarianism. *Social Forces* 89, 3 (2011), 961–982.

[24] Talcott Parsons. 1937. *The Structure of Social Action.* Free Press New York.

[25] Talcott Parsons. 2013. *The social system.* Routledge.

[26] Parvathy Sarat, Prathik Kaundinya, Rohit Mujumdar, and Sahith Dambekodi. 2020. Can Machines Detect if you’re a Jerk? https://rohitmujumdar.github.io/projects/aita.pdf

[27] John Finley Scott. 1971. *Internalization of norms: A sociological theory of moral commitment.* Prentice-Hall.

[28] Richard A Settersten Jr and Gunhild O Hägestad. 1996. What’s the latest? Cultural age deadlines for family transitions. *The Gerontologist* 36, 2 (1996), 178–188.

[29] Henri Tajfel. 1973. The roots of prejudice: Cognitive aspects. *Psychology and race* (1973), 76–95.

[30] Lisa K Tamres, Denise Janicki, and Vicki S Helgeson. 2002. Sex differences in coping behavior: A meta-analytic review and an examination of relative coping. *Personality and social psychology review* 6, 1 (2002), 2–30.

[31] Mike Thelwall and Emma Stuart. 2019. She’s Reddit: A source of statistically significant gendered interest information? *Information processing & management* 56, 4 (2019), 1543–1558.

[32] Nick Yee, Jeremy N Bailenson, Mark Urbanek, Francis Chang, and Dan Merget. 2007. The unbearable likeness of being digital: The persistence of nonverbal social norms in online virtual environments. *CyberPsychology & Behavior* 10, 1 (2007), 115–121.

[33] Karen Zhou, Ana Smith, and Lillian Lee. 2021. Assessing Cognitive Linguistic Influences in the Assignment of Blame. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media.* 61–69.

A REGULAR EXPRESSION

In the following we report the regular expression we developed in order to extract gender and age information from titles of submissions in r/AITA. It uses the Python syntax. Line breaks are added for readability.

```
(?i)\b((I[\s']*(?:a?m|ve|have)?(\san?)?)|
(my)|(myself)|(me))\b[\s\,\:]+(([\(\[\s]*
((?P<age1>[0-9]{2})[\s\,]*(?P<gender1>[mf]))
[\s\,\.]*[\)\]
\s]+)|(([\(\[\s]*(?P<gender2>[mf]))[\s\,\.]*
(?P<age2>[0-9]{2})[\s\,\.]*[\)\]\s]+))
```

To aid the interpretation of this regular expression, Figure 7 shows its corresponding state diagram, obtained from debuggex. The authors wish to thank Lorenzo Betti for his help in improving this regular expression.