Can Anyone Become an Internet Troll?

Jure Leskovec (@jure)

Including joint work with Justin Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil
Two Metaphors for the Web

Web as a library  Web as a social place

You?  Me!

No, me!
Web as a Social Place

Crucial for the success of the Web is the ability of users to contribute content and provide feedback.
We worried that Byrd was gonna get lonely set up a cat cam, this is what we saw

Our new kitten has officially stolen my heart

Posted about Manfred earlier this week. He

Reddit, meet Hercules. This is what he does in the bathroom.

My Best Little Buddy. :)
Can C++ code be valid in both C++03 and C++11 but

Is it possible for C++ code to conform to both the C++03 standard and the C++11 standard? Different things depending on under which standard it is being compiled?

13 I'm pretty sure auto could result in a situation like this – OMGtechy Apr 13 at 19:15

6 Yes. One example is `>>` when used in a template. You can come up with a situation for both standards. Another one that I'm sure would be easy to find changes for is... Apr 13 at 19:15

4 Here's a nice article on the `>>` situation: gustedt.wordpress.com/2013/12/15/... –

5 @OMGtechy: I don't think auto can cause this. With the old meaning, an auto type name; with the new meaning, a type name is not permitted. – Keith Thompson

2 Closing this as too broad would be a shame, this is question likely to be asked much good to have a solid canonical question to point to that is not closed. – Shafik Y
Facebook privacy now defaults to friends only

By Doug Gross, CNN
updated 3:39 PM EDT, Thu May 22, 2014 | Filed under: Social Media

Hi Charlie! Before you post, we wanted to make sure you’re all set.

On Thursday Facebook bowed to pressure and said it would reduce the default settings for new users to be visible to friends only, not everyone on Facebook.

The social networking giant released its plans last year to go with a “broadly social” approach — making users’ information visible to everyone by default. But Facebook has since scrapped that plan and now says that new users will post to a closet of friends only.

Facebook said recently that the change was designed to help new users start with a private “core” and then gradually share more information with the world. Facebook has made previous changes to privacy with the goal of helping users better manage their personal information.

In the past, Facebook users had the choice to set their profile to “public,” making their posts and information visible to anyone online, or to leave it at “friends only,” which means only other Facebook users can see their profile and posts.

Facebook has been under pressure from users to make privacy a priority after a series of high-profile data breaches and privacy concerns.

Widow hugs husband’s killer in court

More from CNN Video:

Serial killer’s daughter recalls horrors

35 Comments

Join the discussion...

Tom 7 hours ago
If you're posting something to facebook, it shouldn't be anything you wouldn't print and tape to the front door of a local grocery store.
Two Metaphors for the Web

Web as a library
- One static version
- Everyone’s experience is fixed
- No user interactions

Web as a social place
- Complex dynamic feedback effects
- One user’s experience is a function of other users’ experiences
- User interactions
Challenges

Want users to engage and contribute (good) content (and make others’ experiences better)

- How do we keep users engaged?
- How do we promote good behavior?
- How do we ensure health of the community and the content ecosystem?
Challenges

- We know a lot about overall average user behavior in online communities.
- But many times few users can have a disproportionate (negative) effect on the entire community.
- How do we study and model behavior of such users?
- Have we use the insights to improve online communities?
Antisocial Behavior

Internet Troll is a person who sows discord on the Internet

- by starting arguments and upsetting people
- by posting inflammatory, extraneous, or off-topic messages in an online community

[wikipedia]
Facebook: The encyclopedia of beauty?

By Amanda Enayati, Special to CNN

Updated 9:13 PM ET. Fri March 16, 2012

Demi Lovato and Miley Cyrus have used social media to send positive messages about body image.

One day last year, Amanda Coleman decided to quit Facebook.

It wasn’t one precipitating event that led her to her decision, but a slow build -- a series of disturbing, sometimes anguished conversations.

Coleman, a college student and the president of her sorority, found herself spending a lot of time counselling young girls, many of them freshman at the university.

Perceptions of Beauty

Dieting companies target men

Men face the issue at the same rate as women, but until recently, ads for weight loss products mostly featured female faces and voices.

Facebook: Encyclopedia of beauty?

Because of social networks, the field of competition has expanded dramatically. Now you’re competing with the best pictures and the ebullient status updates of every girl you know.

Fat: New ugly on the playground

Children as young as 3 worry about being fat. An obsession with skinny ideals is trickling down to our youngest generation.

Sex, lies and media

There’s a new wave of activists challenging our society’s notions of beauty.
Girls are focused on their image way too much. This isn't new. And when they talk about other girls 'fat', 'ugly', and 'b*tch' are recurring words.
It also shows that Islam and Christianity teaching women to dress modest could be right after all.
It also shows that Islam and Christianity teaching women to dress modest could be right after all.

Religious nut alert

Clearly that is the only logical conclusion to this article. Now if you'll excuse me, I need to iron my tarp. I have work on Monday, and I want to appear 'modest'.

fail at life. go bomb yourself.
Both of these little skanks are ugly to the bone.
Both of these little skanks are ugly to the bone.
PROBLEM: Trolls disrupt online discussions

Facebook is for losers with no friends in real life.

Less positive
Soccer star Muamba still 'critical' after on-pitch collapse

Suck a f****ng c@ck u d@uchebag

More profanity
Ex-NBA star returns to inner city, brings hoop dreams

*sniff* This had me tearing up in the office. Great job Penny! You were always a class act on and off the court. Your Grandmother is applauding you again in heaven!

Fantastic story. Kudos to Penny Hardaway. This is what you call giving back.

Penny, once again you show why you are one of the best of the league. Always a class-act. Always a good example.
Penny, once again you show why you are one of the best of the league. Always a class-act. Always a good example.

I have a question for CNN.......why as a reader do I not see any articles similar to this about white NBA basketball players?......every single touchy feely story is about a black ball player and then all you do is blast Tim Tebow.........YOU GUYS MAKE ME SICK AS A READER!

Off-topic: Less similar to previous posts

How many white NBA players grew up...
many stories of charitable efforts by white NBA players, but this isn't a story about a Black NBA player...it's a story about someone who came back to their roots to contribute.

\[ b\textbackslash c \text{ young black men need to see examples from their own race. They need to see that even minorities can succeed and give back instead. They connect with people of their own race. It allows for mentorship and guidance.} \]

A story about a millionaire helping kids in his poor neighborhood personally, and being a positive role model makes you sick as a reader. Gotta love conservatives.
Discussion Communities

Complete data over 18 months, 
~1.7 million users generated 
~40 million posts and 
~100 million votes
How Common is Bad Behavior?

<table>
<thead>
<tr>
<th></th>
<th>CNN</th>
<th>IGN</th>
<th>BREITBART</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post Deletion Rate</td>
<td>21.4%</td>
<td>2.3%</td>
<td>2.7%</td>
</tr>
<tr>
<td>User Ban Rate</td>
<td>3.3%</td>
<td>1.7%</td>
<td>2.2%</td>
</tr>
</tbody>
</table>
This Talk: Trolling

- Large-scale investigation of trolling in online discussion communities

- Common theme: Complementing observational data with experiments
1) Who are trolls?
   - Mechanisms behind trolling behaviors: Mood and discussion context

2) Does feedback mitigate trolling?
   - Votes lead to cascading of bad behaviors

3) Can we detect trolling early?
   - Predicting whether a given user will be blocked from the community
Who are these people?

What makes them troll?

Why is trolling so common?
Studies on Trolling

Prior research suggests:

- Trolls are born and not made  
  [Baker ‘01; Herring et al. ‘11; Shachaf-Hara ‘10]
- Trolls have particular personality  
  [Buckels et al. ’14; Raine ‘02]
- Trolls are sociopathic individuals  
  [Rensin ’14; Schwartz ‘08]
- Q: Are trolls really special people or can anyone become a troll?
Confessions of a former internet troll

by Emmett Rensin on December 16, 2014

Internet Trolls Really Are Horrible People

Behind the online comments: the psychology of internet trolls
Hypotheses: Mood & Context

**H1:** Negative mood increases a user’s likelihood of trolling

**H2:** The discussion context affects a user’s susceptibility to troll.
Designing an Experiment

MTukers only care about being paid.

How do we design an experiment?
Designing an Experiment

Two-by-two experimental design:

- **Step 1: Mood – Take a test (5min)**
  - Easy test & tell people they did well
  - Hard test & tell people they did poorly

- **Step 2: Context – Be in a discussion**
  - Neutral context of neutral comments
  - Negative context of negative comments

People do not know the real purpose. Are debriefed at the end.
Controlling Mood: PosMood

Qualification Test

Instructions

- Below is a series of simple questions, many of which you should be able to answer correctly.
- You will have **five minutes** to complete all the questions.
- Currently, average performance is **8 or more correct answers**.
- You are allowed to use pen and paper, but not any electronic aids (including the Internet).
- Your performance on this task will not affect your payment on the task.

Unscramble the following letters to form an English word:
"P A P H Y"

Type in your answer.

Subtract three thousand from five thousand. Write your answer in words.

Type in your answer.

How many 'l's are there in these sentences?

266.5 seconds left
Controlling Mood: NEGMOOD

Qualification Test

Instructions

- Below is a series of simple questions, many of which you should be able to answer correctly.
- You will have **five minutes** to complete all the questions.
- Currently, average performance is **8 or more correct answers**.
- You are allowed to use pen and paper, but not any electronic aids (including the Internet).
- Your performance on this task will not affect your payment on the task.

Unscramble the following letters to form an English word:

"D E A N Y O N"

Type in your answer.

Subtract six thousand, seven hundred eighty-three from eight thousand, eleven. Write your answer in words.

Type in your answer.

How many 'I's are there in these sentences?
News of the Day

I’m Voting for Hillary Because of My Daughter

Back in the 2008 primary season, I supported Hillary Clinton. That choice wasn’t easy for me, especially as Barack Obama advanced and became ever more compelling as a candidate. I felt spoiled as the year progressed: two great,
Context: NEGCONTEXT

News of the Day

I’m Voting for Hillary Because of My Daughter

Back in the 2008 primary season, I supported Hillary Clinton. That choice wasn’t easy for me, especially as Barack Obama advanced and became ever more compelling as a candidate. I felt spoiled as the year progressed: two great,

Top Comments Sorted by Best

User1337 · 9 weeks ago
Oh yes. By all means, vote for a Wall Street sellout -- a lying, abuse-enabling, soon-to-be felon as our next President. And do it for your daughter. You’re quite the role model.
1 · Reply

User9054 · 9 weeks ago
Hillary is a [censored]. I am voting with my [censored] for Putin. /s
0 ·

Write a comment...

Post Comment
Mechanical Turk Experiment

667 participants, 40% female, age 34.2, 54% Democrat, 25% Moderate, 21% Republican
Contributed 791 posts (37.8 words/post), 1392 votes
Results: Test Performance

- **PosMood** (easy test): 11 out of 15 questions correct
  - Outperforming stated avg. score of 8

- **NegMood** (hard test): 1.9 out of 15 questions correct
  - After below stated avg. score of 8

Does the test score affect user’s mood?
Results: Does it affect mood?

- After the test: 65 questions on how people were feeling
  - Based on the Profile of Mood States

- People in \textsc{negmood} felt worse:
  - Higher anger, confusion, depression, fatigue, and tension scores

- How good were the comments?
## Comments: MOOD

<table>
<thead>
<tr>
<th>NeutralContext</th>
<th>Prop. Troll Posts</th>
<th>NegContext</th>
<th>Prop. Troll Posts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>PosMood</strong></td>
<td><strong>NegMood</strong></td>
<td></td>
</tr>
<tr>
<td>NeutralContext</td>
<td>0.35</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>NegContext</td>
<td>0.47</td>
<td>0.68</td>
<td></td>
</tr>
</tbody>
</table>

- Users in the **negative mood** condition are more likely to troll (~80% higher odds, p < 0.05)
### Comments: CONTEXT

<table>
<thead>
<tr>
<th></th>
<th>Prop. Troll Posts</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PosMood</td>
<td>NegMood</td>
</tr>
<tr>
<td>NeutralContext</td>
<td>0.35</td>
<td>0.49</td>
</tr>
<tr>
<td>NegContext</td>
<td>0.47</td>
<td>0.68</td>
</tr>
</tbody>
</table>

- Users in the **negative context** condition are more likely to troll (~60% higher odds, p < 0.05)
Negative mood and context lead to more negative posts
Bad mood and context make people write bad posts.

How do we combat that?
Sites allow users to vote on the content and hopefully bad posts get voted down.
We worried that Byrd was gonna get lonely, so we set up a cat cam, this is what we saw

Our new kitten has officially stolen my heart.

Posted about Manfred earlier this week. He seems to have a soft spot for the bathroom.

Reddit, meet Hercules. This is what he does.

My Best Little Buddy. :)
Can C++ code be valid in both C++03 and C++11 but

Is it possible for C++ code to conform to both the C++03 standard and the C++11 standard? It seems like different things depending on under which standard it is being compiled.

c++  c++11  language-lawyer  c++03

share | improve this question

edited Apr 17 at 20:02

Angew
39.5k 5 41 88

OMGtechy Apr 13 at 19:15

I'm pretty sure auto could result in a situation like this...

Yes. One example is >> when used in a template. You can come up with a situation for both standards. Another one that I'm sure would be easy to find changes for is

OMGtechy Apr 13 at 19:15

Here's a nice article on the >> situation: gustedt.wordpress.com/2013/12/15/

@OMGtechy: I don't think auto can cause this. With the old meaning, an auto type name; with the new meaning, a type name is not permitted.

Closing this as too broad would be a shame, this is question likely to be asked in good to have a solid canonical question to point to that is not closed.
Join the discussion...

Tom · 7 hours ago
If you're posting something to facebook, it shouldn't be anything you wouldn't print and tape to the front door of a local grocery store.

ccw101 ➔ Tom · 7 hours ago
I hate Facebook for the fact the only person you have control over is yourself. I have seen full grown adults get angry at their own children and rip them a new one on their Facebook home page!
If adults can be so ST***id then what do kids do?
Facebook is scary. And has given people the opportunity to use it to cause home break in's, ruined reputations, fights, suicides etc.

IAmNotATroll ➔ ccw101 · 7 hours ago
Come now. I thoroughly enjoy watching my in-laws publicly argue and shred each other to pieces over Facebook.

Furby ➔ IAmNotATroll · 6 hours ago
I had a distant cousin try to blackmail my mom on FB publically. Me and her became real close after that - and not in the way you want to get close to someone. Some people are just plain dumb
When a user evaluates a comment, she is indirectly evaluating the author.

Do such evaluations help authors contribute better content?
How to Summarize Feedback?

- Number of up-votes?
- Up-votes minus Down-votes?
- Fraction of up-votes?
P: Up-Votes

Number of up-votes:

| 15+ | 0- | vs | 15+ | 15- |

Doesn’t account for down-votes
P–N : Up – Down Votes

Difference in up-votes and down-votes:

5+ 0− vs 50+ 45−

Doesn’t account for proportion of up-votes
P/(P+N): Fraction Upvotes

Proportion of up-votes:

| 4+ 1- | vs | 40+ 10- |

Doesn’t account for total number of votes
How is Feedback Perceived?

Measure how users perceive votes

- Crowdsourcing task:
  Workers rate, on a 7 point scale, if they felt more positive, or negative about receiving a certain number of up/down-votes.
Results: Perceiving Votes

User ratings are independent of the total number of votes.
<table>
<thead>
<tr>
<th></th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>P</strong></td>
<td>0.410</td>
</tr>
<tr>
<td>Number of up-votes</td>
<td></td>
</tr>
<tr>
<td><strong>P-N</strong></td>
<td>0.879</td>
</tr>
<tr>
<td>Difference in up/down-votes</td>
<td></td>
</tr>
<tr>
<td><strong>$P/(P+N)$</strong></td>
<td>0.920</td>
</tr>
<tr>
<td>Proportion of up-votes</td>
<td></td>
</tr>
</tbody>
</table>
Evaluation of a Post

- A post is positively/negatively evaluated when the proportion of up-votes is higher/lower than a given threshold.

\[ \frac{P}{P+N} = \frac{9}{9+1} = 0.9 \geq \text{High Threshold} \]
How does positive/negative feedback influence subsequent behavior?

defined by the proportion of up-votes
Effects of Evaluations

- Post quality
  - How well you write
- Community bias
  - How people evaluate you
- Posting frequency
  - How regularly you post
- Voting behavior
  - How you vote on others
Do Users Improve?

Operant conditioning predicts that feedback would guide authors towards better behavior:

- up-votes are “reward” stimuli
- down-votes are “punishment” stimuli

Or Do They Get Worse?

Feedback can have negative effects:

- People given only positive feedback tend to become complacent
- Also, bad impressions are quicker to form and more resistant to disconfirmation

Matching Similar Users

Compare pairs of users who were evaluated differently on similar content

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects.
Details of Matching

Match pairs of users where one got positive and one got negative evaluation

- Match based on
  - history text quality
  - number of posts
  - post length
  - overall proportion of up-votes

Text quality determined by training a machine learning model using text features, validated using crowd workers.
Results: Neg-Votes Increase

Before

After

Proportion of up-votes

Positive evaluation

Negative evaluation

Results: Neg-Votes Increase
Content vs. Community

Can valuations be explained by textual or community effects?

- Down-voting because of post content
  - I hate ...

- Down-voting because the community dislikes the author
  - We don’t like you.
Results: Negativity Bias

- Text quality drops significantly after a negative evaluation but does not change after a positive evaluation

p < 0.05 in all communities

To learn more about these types of effects, see Kanouse, D. E., & Hanson Jr, L. R. (1987). Negativity in evaluations.
Evaluations can Affect...

...Community Bias

(How people perceive you)

- How does community perception of a user change after an evaluation?
<table>
<thead>
<tr>
<th>Actual Evaluation $P/(P+N)$</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Quality</td>
<td>0.8</td>
</tr>
<tr>
<td>Community Bias</td>
<td>0.9-0.8</td>
</tr>
<tr>
<td></td>
<td>$= +0.1$</td>
</tr>
</tbody>
</table>
Compare community bias before and after the evaluation.
Communities Discriminate

Posts made after a negative evaluation were perceived worse than those made after a positive evaluation

- On CNN, on average, your posts are being evaluated 48% worse than would be predicted by text quality

\[ p < 0.05 \text{ in all communities} \]
Summary of Effects

- Positive Eval.

- Similar History

- Similar Text Quality

- Worse Text

- Worse Perception

Before

Negative Eval.

After

More Positive
Evaluations can Affect...

...Posting Frequency

(How often you post)

- Does feedback regulate posting quantity?
Feedback and Posting Activity

- Do users post more after a positive/negative evaluation?
Negativity Increases Activity

Users who receive negative feedback post more frequently
Evaluations can Affect...

...Voting Behavior

(How you vote on others)

- Does feedback result in subsequent backlash of negative behavior?
More Down-Votes Given

Users who receive negative feedback are more likely to down-vote others.
Summary so far...

- Negatively-evaluated users
  - Write worse (and more!),
  - Are evaluated worse by the community
  - And evaluate others worse

- Positively-evaluated users
  don’t do any better
Is there a downward spiral in communities?
Downward Spiral in CNN

Proportion of down-votes is increasing over time

- Jan: 0.16
- Mar: 0.18
- May: 0.2
- Jul: 0.24

0.8m down-votes
1.7m down-votes
WHY WE'RE SHUTTING OFF OUR COMMENTS

STARTING TODAY, POPULARSCIENCE.COM WILL NO LONGER ACCEPT COMMENTS ON NEW ARTICLES. HERE'S WHY.

By Suzanne LaBarre  Posted September 24, 2013

We're turning comments off for a bit

by Joshua Topolsky  I I February 2nd 2010 at 11:25 am

Pewdiepie Says Turning Off Comments Made Him Happier

Patricia Hernandez  Filed to: PEWDiePIE  11/05/14 3:30pm
Can we identify antisocial behavior quickly and prevent downward spirals?

Antisocial Behavior in Online Discussion Communities
AAAI ICWSM, 2015.
Prediction Task

- **Task:** Detect which users will be banned in the future

- Post: # of words
  - ...  

- Activity: posts/day
  - ...  

- Community: % upvoted
  - ...  

- Moderator: deletions
  - ...  

- ~19k users CNN, balanced dataset, first 10 posts
Prediction Results

<table>
<thead>
<tr>
<th>Method</th>
<th>ROC AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag of Words</td>
<td>0.70</td>
</tr>
<tr>
<td>Deletion Rate</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Bar chart showing the ROC AUC for Bag of Words and Deletion Rate with values 0.70 and 0.74 respectively.
Prediction Results

<table>
<thead>
<tr>
<th></th>
<th>ROC AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag of Words</td>
<td>0.70</td>
</tr>
<tr>
<td>Deletion Rate</td>
<td>0.74</td>
</tr>
<tr>
<td>Post</td>
<td>0.62</td>
</tr>
</tbody>
</table>

- **Bag of Words**: 0.70
- **Deletion Rate**: 0.74
- **Post**: 0.62
Prediction Results

<table>
<thead>
<tr>
<th>Feature</th>
<th>ROC AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag of Words</td>
<td>0.70</td>
</tr>
<tr>
<td>Deletion Rate</td>
<td>0.74</td>
</tr>
<tr>
<td>Post</td>
<td>0.62</td>
</tr>
<tr>
<td>+ Activity</td>
<td>0.73</td>
</tr>
<tr>
<td>+ Community</td>
<td>0.83</td>
</tr>
<tr>
<td>+ Moderator</td>
<td>0.84</td>
</tr>
</tbody>
</table>

0.5  0.6  0.7  0.8  0.9
Time to Detection

User lifetime:

- Red: >= N
- Blue: 10–20
- Purple: 40–80
- Orange: 5–10
- Green: 20–40
- Yellow: 80+

AUC

Using First N Posts

- AUC values for different user lifetimes are shown in the graph.
- The graph illustrates how the AUC changes with different numbers of posts used for prediction.
- The performance generally increases as more posts are used, with the AUC approaching 0.90 for longer user lifetimes.

The figure also shows how classifier performance changes with different time periods. For instance, performance seems to peak near ten posts (Figure 7b), with the most performant feature being the proportion of up-votes received (0.75), followed by the number of posts per post deletion (0.73). Community features were next, with the most performant feature being the proportion of posts deleted (0.76). Moderator features are the strongest predictors of being subsequently banned (individual feature set AUC=0.75 for moderator features).

Future study could focus on the evolution of antisocial behavior and how predictive models can be improved. This paper presents a data-driven study of antisocial behavior in online discussion communities by analyzing users that will eventually be banned not only write worse words but also more posts over time, but the community becomes less tolerant of them. Next, it proposes a typology of antisocial users based on post deletion rates. Finally, it introduces a system for identifying undesired users early on in their community life.

How generalizable are these classifiers?

This experiment aimed to determine how generalizable the classifiers are to new domains. A data-driven approach was used, involving the use of post features such as the proportion of deleted posts, as well as the number of posts and up-votes received. The approach has several limitations, such as the lack of fine-grained labels for more covert instances of antisocial behavior.

The results showed that the classifiers performed well in cross-domain experiments, with mean AUC=0.79 across all communities. However, the generalizability of these features, as well as the generalizability of the models learned on single communities, suggests a need for further investigation.

This paper highlights the importance of characterizing the behavior of antisocial users and the generalizability of predictive models. Future work could focus on developing more fine-grained labels for more covert instances of antisocial behavior and improving the generalizability of predictive models.
Results: Cross-Domain

<table>
<thead>
<tr>
<th>Tested on</th>
<th>CNN</th>
<th>IGN</th>
<th>BREITBART</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROC AUC</td>
<td>0.84</td>
<td>0.74</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>0.69</td>
<td>0.83</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>0.74</td>
<td>0.75</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Trained on
Conclusion

- Web is a social place!
  But people are less inhibited online
- Important to understand interaction mechanisms and behavioral patterns
- Towards healthier and more welcoming online communities
Summary

Large-scale longitudinal analysis of antisocial behavior in discussion forums

Combination of data analysis and human experiments

Trolls are not “special” people

- Mood and discussion context increase the probability of trolling
Summary

Bad behavior can spread

- The effects of negative feedback are more pronounced
  - Write worse and more
  - Are evaluated worse by others
  - Evaluate others worse

- Positively-evaluated users, on the other hand, don’t do any better
Summary

- **Prediction task:** Predicting whether a user will be banned in the future
- After observing just 10 posts, we can accurately predict whether the user will be banned in the future
- Generalization performance across communities
Further Questions

- Social feedback loops: Are trolls socially created by harsh feedback?
- Reputation: Effects of user reputation
- Role of moderators: Are moderators effectively helping and moderating
- Linguistic norms: Linguistic analysis
- Feedback mechanism: Communities with only positive feedback (likes)
KEEP CALM AND DON'T FEED THE TROLL
References

- **How Community Feedback Shapes User Behavior**