

Forum:[Hack Forums](#)

Thread: [Scoring toxicity levels of posts w/ machine learning - Useful feature for interaction](#)

W?lr?? - June 22, 2017, 11:59 pm

I believe that this would be very useful to Hackforums, since I often see people who post things attacking or being negative to each other. I found an API a few weeks ago that basically scores the "potential impact of your writing" with 0% being not toxic and 100% being toxic.

Check it out here: <https://www.perspectiveapi.com>

From my testing, it is right basically every time. It would be cool if the API worked with this site and under every post it had a score saying the % score.

Examples

Vitani - June 23, 2017, 12:08 am

While the perspectiveapi project itself is a pretty cool thing, I don't think this is necessary as the report system gets most if not all posts who violate HF policy already.

Aside from that, It's also not allowed to make moderator or moderation suggestions. Reference: [Suggestions That Will NOT Be Accepted](#)

W?lr?? - June 23, 2017, 12:13 am

(Jun 23, 2017 12:08 AM)Vitani Wrote: [\[CITE\]](#) While the perspectiveapi project itself is a pretty cool thing, I don't think this is necessary as the report system gets most if not all posts who violate HF policy already.

Aside from that, It's also not allowed to make moderator or moderation suggestions. Reference: [Suggestions That Will NOT Be Accepted](#)

Merely the intention of it could be for cosmetic use and for those who post something to re-think before posting something that can be interpreted as toxic. It does have some moderation benefits, but as the rules have stated, that would not be useful from the perspective of a regular user, so I can not make suggestions for that aspect.

I think it's one more step forward in technology and something that forums will have in the future. With Hackforums being a tech forum, I see it being a great part of interacting and posting throughout the forum.

Vitani - June 23, 2017, 12:21 am

(Jun 23, 2017 12:13 AM)W?lr?? Wrote: [\[CITE\]](#) Merely the intention of it could be for cosmetic use and for those who post something to re-think before posting something that can be interpreted as toxic. It does have some moderation benefits, but as the rules have stated, that would not be useful from the perspective of a regular user, so I can not make suggestions for that aspect.

I think it's one more step forward in technology and something that forums will have in the future. With Hackforums being a tech forum, I see it being a great part of interacting and posting throughout the forum.

It's definitely not a bad idea and all in all it's a pretty cool project. I am not sure if this will actually make the LQ posters rethink their posts but if possible it's worth a try.

I do think that this system could be fooled and people could bypass it by posting random words. I'll look into it more to be sure of this though.

As for moderation, I still think it's not necessary but as you said, for users itself it could have some benefits.

Going to look into perspectiveAPI more to actually see how it works and post back here.

W?lr?? - June 23, 2017, 12:28 am

The biggest issue that I've seen is that it is very sensitive. It works well for formal writing but I am not quite sure how well it would work in informal writing. Whenever I use vulgar language, it immediately increases it by a lot.

Check this example. Under Hackforums rules, this would technically be allowed. However, it triggers it to 100% because this a very low quality post when writing formally. This isn't an issue because it helps keep the forums clean and high quality, even though technically this message isn't against the rules. If this user were to explain what he means more reasonably, then it would have not triggered a toxic warning.

Omniscient - June 23, 2017, 1:00 am

I don't think I'd do this in the phase it's currently at. I'll look more into it though.

Picturesque - June 23, 2017, 1:03 am

Not the best system.

Jendo - June 23, 2017, 1:07 am

(Jun 23, 2017 01:03 AM)SellingForums Wrote: [\[CITE\]](#)

Not the best system.

Imafooo

the system is not smart to understand the meaning of the sentences

it just has a list of toxic words, it checks if the sentence has any of them.

Picturesque - June 23, 2017, 1:08 am

(Jun 23, 2017 01:07 AM)Jendo Wrote: [\[CITE\]](#) Imafooo

the system is not smart to understand the meaning of the sentences

it just has a list of toxic words, it checks if the sentence has any of them.

Exactly. Throw a few of the words "please", "like", and use correct grammar and it doesn't flag it as being toxic.

It marked "I'm planning on shooting up my school tomorrow morning." as only 19% toxicity.

Vitani - June 23, 2017, 1:34 am

Just looked into it a bit more, the system is largely based of grammar and a few specific "lq" words.

It has potential but for now I don't think it would have enough benefits to add this.

Good suggestion though!

TB12 - June 23, 2017, 1:56 am

Once it's more developed I think this would be a cool feature. Shouldn't be added for now though

vaiperboy - June 23, 2017, 3:15 am

so what if someone's toxic? who tf cares :|

Lil Uzi Vert- - June 23, 2017, 5:57 am

(Jun 23, 2017 03:15 AM)vaiperboy Wrote: [\[CITE\]](#) so what if someone's toxic? who tf cares :|
I do

I don't want my feelings to be hurt :(

M?xim - June 23, 2017, 8:46 am

In your example, that is pretty much just an opinion on the government. Doesn't make it toxic...
Really no need for this lol.

W?lr?? - June 23, 2017, 9:31 am

(Jun 23, 2017 08:46 AM)Maxim Wrote: [\[CITE\]](#) In your example, that is pretty much just an opinion on the government. Doesn't make it toxic...
Really no need for this lol.

Toxic levels are mainly measured by how they impact the conversation. A toxic level of 0%-10% is likely to be very beneficial and resourceful while a toxic level of 70%-100% means it is not contributing a valuable point to the conversation and possibly being negative to the conversation.

Your message got 34% likely to be perceived as "toxic" because it is not beneficial to the conversation. You are not displaying powerful points in your message. The message almost ridicules those who have a different opinion. This is something that you could debate, however generally people read comments like those as "arrogant sounding". The "..." and "lol" establish the message that:

- You believe the idea is stupid and that the example I posted was not measuring toxicity levels
- You believe that anyone who doesn't agree with your point of view is stupid and ridiculous.

M?xim - June 23, 2017, 9:34 am

(Jun 23, 2017 09:31 AM)W?lr?? Wrote: [\[CITE\]](#) Toxic levels are mainly measured by how they impact the conversation. A toxic level of 0%-10% is likely to be very beneficial and resourceful while a toxic level of 70%-100% means it is not contributing a valuable point to the conversation and possibly being negative to the conversation.

Your message got 34% likely to be perceived as "toxic" because it is not beneficial to the conversation. You are not displaying powerful points in your message. The message almost ridicules those who have a different opinion. This is something that you could debate, however generally people read comments like those as "arrogant sounding"

Fuck me this suggestion is stupid. How is my toxic level now?

W?lr?? - June 23, 2017, 9:36 am

(Jun 23, 2017 01:03 AM)SellingForums Wrote: [\[CITE\]](#)

Not the best system.

PerspectiveAPI measures how things impact a conversation, so this is not impacting the conversation negatively. You are displaying a point proficiently without offending people of another perspective. And keep in mind that some places have the age of consent at 16, so your point has been proven invalid.

(Jun 23, 2017 09:34 AM)Maxim Wrote: [\[CITE\]](#) Fuck me this suggestion is stupid. How is my toxic level now?

Obviously a 100% because it is not affecting the conversation positively.

M?xim - June 23, 2017, 9:42 am

(Jun 23, 2017 09:36 AM)W?lr?? Wrote: [\[CITE\]](#) Obviously a 100% because it is not affecting the conversation positively.

Just wanted to test.

Welp, I think that it shouldn't be implanted now at least, doesn't seem to be advanced or smart enough at this stage.

W?lr?? - June 23, 2017, 9:47 am

(Jun 23, 2017 09:42 AM)Maxim Wrote: [\[CITE\]](#) Just wanted to test.

Welp, I think that it shouldn't be implanted now at least, doesn't seem to be advanced or smart enough at this stage.

Yes it was only released to the public since 4 months ago. It's quite difficult to get training data, but I assume once they start partnering up with big social media/forum sites, than it will start becoming very proficient in understanding. Until then, I'll let this suggestion sit for a while and maybe in a few months we can come back to it

Charger - June 23, 2017, 4:07 pm

This is a forum, not a place to submit your essay.

Top 15 useless suggestions I've ever seen.

TyrantKingBen - June 23, 2017, 4:25 pm

Text analytics is cool but I don't see the usefulness for this to be on the front-end whatsoever. However, Omni could train a model to classify a shitpost and implement that in the back-end to help remove some spamming. I don't think it's worth the time or effort though.

X X X X - June 23, 2017, 4:46 pm

Interesting but I don't really like having it under every post. Maybe only visible for staff?

jujuger - June 23, 2017, 5:01 pm

(Jun 23, 2017 01:07 AM)Jendo Wrote: [\[CITE\]](#) Imafooo

the system is not smart to understand the meaning of the sentences

it just has a list of toxic words, it checks if the sentence has any of them.

This. It doesn't under the blanket term of "machine learning", when it is obvious that the neural network it uses simply concluded "lol okay so if it contains words from database "negative", then it is bad".

It will not be able to detect sarcasm probably.

In fact, this is probably the dumbest idea I have ever heard. Why what if someone says something containing negative words that's IMPORTANT AND USEFUL, such as "warning: hackforums is

currently under a ddos attack, please do not log in"

Or what if I am exposing a scammer on hf? "warning: Omni is a criminal and has scammed \$9000 out of members with his ebook!"

Guys, sometimes "negativity" is useful depending on the social context.

In addition, why would i need a computer to tell me if something is negative? i can see clearly that "kill all jews" is negative and i do not need a computer to tell me this

Otis???????????? - June 23, 2017, 5:41 pm

(Jun 23, 2017 01:03 AM)SellingForums Wrote: [\[CITE\]](#)

Not the best system.

I bursted out laughing when I saw this...I'm horrible lol

ra1d - June 23, 2017, 5:51 pm

What is this??? FamilyFriendlyForums.net?

Shit suggestion imo

+Cloud - June 23, 2017, 5:53 pm

(Jun 23, 2017 09:36 AM)W?lr?? Wrote: [\[CITE\]](#) PerspectiveAPI measures how things impact a conversation, so this is not impacting the conversation negatively. You are displaying a point proficiently without offending people of another perspective. And keep in mind that some places have the age of consent at 16.

Age of consent here is 16 so reading that was fucked for me it pretty much said i wanna have sex with someone under 16, your point is invalid

W?lr?? - June 23, 2017, 6:05 pm

(Jun 23, 2017 05:53 PM)+Cloud Wrote: [\[CITE\]](#) Age of consent here is 16 so reading that was fucked for me it pretty much said i wanna have sex with someone under 16, your point is invalid

It really is merely impossible for an algorithm to judge the opinion in a message and whether it is "messed up" according to society. It would take years before algorithms could inherit those abilities.

My main vision with Hackforums is for those who are being an asshole to realize that they are being toxic. Although it wouldn't do anything meaningful, it's a step in the direction of helping users post more HQ. Many users have been so LQ that Omniscient has banned them just based on one reply. Most, if not all, would have been rated a 100% using the PerspectiveAPI. I am not defending the system yet, as it clearly needs more work, however it should be something to consider in the future.

(Jun 23, 2017 05:01 PM)jujuger Wrote: [\[CITE\]](#) Guys, sometimes "negativity" is useful depending on the social context.

In addition, why would i need a computer to tell me if something is negative? i can see clearly that "kill all jews" is negative and i do not need a computer to tell me this

One problem with negativity is that if used incorrectly, it negatively impacts the conversation as well. PerspectiveAPI has been very good at recognizing when someone is posting something without adding real content value.

(Jun 23, 2017 04:46 PM)X X X X Wrote: [\[CITE\]](#) Interesting but I don't really like having it under every post. Maybe only visible for staff?

From what I have researched, the whole idea is based on theory. How it would be made in application would be up to if and how the idea was implemented.

Also notice many of the people who are against the idea are being toxic themselves.

thales - June 23, 2017, 6:12 pm

I mean... doesn't work well for tech talk.

W?lr?? - June 23, 2017, 6:15 pm

(Jun 23, 2017 06:12 PM)thales Wrote: [\[CITE\]](#) I mean... doesn't work well for tech talk.

Sir... please scroll up and read what I have already said. It would be many years before an algorithm would be able to recognize these sorts of things, but as of now it's a work in progress.

X X X X - June 23, 2017, 6:26 pm

Took me a couple tries but I got this one :D

+Cloud - June 23, 2017, 6:33 pm

People know when they're being toxic, why do we need something to tell us

PropelTheHack - June 23, 2017, 6:51 pm

(Jun 23, 2017 09:36 AM)W?lr?? Wrote: [\[CITE\]](#) PerspectiveAPI measures how things impact a conversation, so this is not impacting the conversation negatively. You are displaying a point proficiently without offending people of another perspective. And keep in mind that some places have the age of consent at 16, so your point has been proven invalid.

That comment can offend many people all over the world regardless of 'age of consent'.

In the USA, it's usually 18 years old, in other parts of the world it's 16.

Either way, depending on which part of the world you may live in, that comment would be considered 'toxic' due to the nature of the comment which strongly suggests the mindset of the person who wrote that comment.

PropelTheHack - June 23, 2017, 7:00 pm

(Jun 23, 2017 06:05 PM)W?lr?? Wrote: [\[CITE\]](#) My main vision with Hackforums is for those who are being an asshole to realize that they are being toxic. Although it wouldn't do anything meaningful, it's a step in the direction of helping users post more HQ. Many users have been so LQ that Omniscient has banned them just based on one reply. Most, if not all, would have been rated a 100% using the PerspectiveAPI. I am not defending the system yet, as it clearly needs more work, however it should be something to consider in the future.

One problem with negativity is that if used incorrectly, it negatively impacts the conversation as well. PerspectiveAPI has been very good at recognizing when someone is posting something without adding real content value.

While your suggestion is a noble idea, PerspectiveAPI has a LONG ways to go before it can be implemented in a forum based website such as Hack Forums.

To create a program that is basically a glorified bot to search hundreds of messages per day to make sure that certain posting criteria is met for the forum in question is a very large task to achieve.

For example, PerspectiveAPI has to be customized for each forum website.

What will work in Hack Forums may not work in my forum based website because I have different posting criteria that Hack Forums does.

Just saying...

W?lr?? - June 23, 2017, 7:00 pm

I think there still needs to be a lot of work done on this API before it can be useful. So for that reason I am closing the thread.