1. **Problem definition:**

Given comment from any platform , our task is to identify if it has any toxic content and classify it to belong to one or more of the following categories . The comments contains foul languages , derogatory remarks. This could lead to spread of hatred , spread of racial slur , tension in communities , attack on individuals . It is a abuse on freedom of speech .

Groups in which comments can be divided in a below way :-

* 1. Toxic
  2. Severely toxic
  3. Obscene
  4. Threat
  5. Insult
  6. Identity hate

The threat of abuse and harassment online means that many people stop expressing themselves and give up on seeking different opinions. Platforms struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user comments.

So there is a requirement to build  multi-headed model that’s capable of detecting different types of of toxicity like threats, obscenity, insults, and identity-based hate.

1. **Literature survey :**

**Instagram:**

Instagram's rolling out some new moderation tools which will use machine learning to detect not only offensive language and spam, but also, comments that are considered discouraging for other users.

As explained by [Instagram](https://blog.instagram.com/post/162395020002/170629-comments):

*"Many of you have told us that toxic comments discourage you from enjoying Instagram and expressing yourself freely. To help, we've developed a filter that will block certain offensive comments on posts and in live video."*

A screenshot of a cell phone

Description automatically generated

As shown in the above screenshot, users will be able to activate the option within their 'Comment Settings'. Once in place, the system will use [Facebook's DeepText AI classification system](https://www.wired.com/story/instagram-launches-ai-system-to-blast-nasty-comments) to detect and eliminate any comments that it deems to be in violation of Instagram's [Community Guidelines](https://help.instagram.com/477434105621119/).

The system has been 'taught', by analyzing a huge range of examples in comments, how to detect such remarks, and utilizes a range of qualifiers to improve it's results.

As explained by [Wired](https://www.wired.com/story/instagram-launches-ai-system-to-blast-nasty-comments):

*"As with spam, the comments are rated based both on a semantic analysis of the text and factors such as the relationship between the commenter and the poster, as well as the commenter's history. Something typed by someone you've never met is more likely to be graded poorly than something typed by a friend."*

The commenter themself will still see the comment on his/her device, reducing their motivation to try again, as they'll assume that comment is visible to all. But that could also lead for some confusing interactions, especially if your comment is blocked erroneously - which will still occur in some cases.

This adds to Instagram's existing comment moderation tools, including the ability to [block specific words, or emoji](https://www.socialmediatoday.com/social-networks/instagram-rolls-out-personalized-comment-filters-curb-trolls-and-abuse), or your choosing, and the option to [switch off comments on posts](https://www.socialmediatoday.com/social-networks/instagram-adds-new-community-safety-tools-including-ability-switch-comments) entirely.

**Google:**

Google’s parent company, Alphabet, says it plans to apply machine learning technology to promote more civil discourse on the internet and make comment sections on sites a little less awful.

Jigsaw, a technology incubator within Alphabet, says it has developed a new tool for web publishers to identify toxic comments that can undermine a civil exchange of ideas. Starting Thursday, publishers can start applying for access to use Jigsaw’s software, called Perspective, without charge.

**Jigsaw Software Perspective API**

**Perspective API** :

<https://www.perspectiveapi.com/#/home>

Perspective API

Perspective is an API that makes it easier to host better conversations. The API uses machine learning models to score the perceived impact a comment might have on a conversation. Developers and publishers can use this score to give realtime feedback to commenters or help moderators do their job, or allow readers to more easily find relevant information, as illustrated in two experiments below. Our first model identifies whether a comment could be perceived as “toxic” to a discussion.

A screenshot of a cell phone

Description automatically generated

**Research:**

A close up of a newspaper

Description automatically generated

**A screenshot of a cell phone

Description automatically generated**

**Wikipedia :**

The[Wikipedia](https://en.wikipedia.org/wiki/Main_Page) community, the free encyclopedia that is built from a model of openly editable content, is notorious for its toxicity. The issue was so bad that the[number of active contributors or editors—those that made one edit per month—had fallen by 40 percent](https://www.technologyreview.com/s/544036/artificial-intelligence-aims-to-make-wikipedia-friendlier-and-better/) during an eight-year period. Even though there’s not one solution to combat this issue, Wikimedia Foundation, the nonprofit that supports Wikipedia, decided to use [artificial intelligence](https://www.bernardmarr.com/default.asp?contentID=963) to learn more about the problem and consider ways to combat it.

**Collaboration with Wikimedia Foundation and Jigsaw to Stop Abusive Comments**

In one effort to stop the trolls, Wikimedia Foundation partnered with Jigsaw (the tech incubator formerly known as Google Ideas) on a research project called Detox using machine learning to flag comments that might be personal attacks. This project is part of Jigsaw’s initiative to build open-source AI tools to help combat harassment on social media platforms and web forums.

**Solution:**

Then, the project team had the algorithm review 63 million English Wikipedia comments posted during a 14-year period between 2001 to 2015 to find patterns in the abusive comments. What they discovered was outlined in the[Ex Machina: Personal Attacks Seen at Scale paper](https://arxiv.org/pdf/1610.08914.pdf):

* 1. More than 80% of all comments characterized as abusive were made by more than 9,000 people who made less than five abusive comments in a year rather than an isolated group of trolls.
  2. Nearly 10% of all attacks were made by just 34 users.
  3. Anonymous users made up 34% of all comments left on Wikipedia.
  4. More than half of the personal attacks are being carried out by registered users although anonymous users were six times more likely to launch personal attacks. (There are 20 times more registered users than anonymous users.)

1. Now that the algorithms have created more clarity about who is contributing to the community’s toxicity, Wikipedia can figure out the best way to combat the negativity. Although human moderation is likely still needed, algorithms can help sort through the comments and flag those that require human involvement.
2. **Sample data**

The below data is taken from Wikipedia comments section over a period from the year 2001 to 2014

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **id** | **comment\_text** | **toxic** | **severe\_toxic** | **obscene** | **threat** | **insult** | **identity\_hate** |
| **0000997932d777bf** | Explanation  Why the edits made under my username Hardcore Metallica Fan were reverted? They weren't vandalisms, just closure on some GAs after I voted at New York Dolls FAC. And please don't remove the template from the talk page since I'm retired now.89.205.38.27 | 0 | 0 | 0 | 0 | 0 | 0 |
| **000103f0d9cfb60f** | D'aww! He matches this background colour I'm seemingly stuck with. Thanks.  (talk) 21:51, January 11, 2016 (UTC) | 0 | 0 | 0 | 0 | 0 | 0 |
| **000113f07ec002fd** | Hey man, I'm really not trying to edit war. It's just that this guy is constantly removing relevant information and talking to me through edits instead of my talk page. He seems to care more about the formatting than the actual info. | 0 | 0 | 0 | 0 | 0 | 0 |
| **0001b41b1c6bb37e** | "  More  I can't make any real suggestions on improvement - I wondered if the section statistics should be later on, or a subsection of ""types of accidents""  -I think the references may need tidying so that they are all in the exact same format ie date format etc. I can do that later on, if no-one else does first - if you have any preferences for formatting style on references or want to do it yourself please let me know.  There appears to be a backlog on articles for review so I guess there may be a delay until a reviewer turns up. It's listed in the relevant form eg Wikipedia:Good\_article\_nominations#Transport  " | 0 | 0 | 0 | 0 | 0 | 0 |
| **0001d958c54c6e35** | You, sir, are my hero. Any chance you remember what page that's on? | 0 | 0 | 0 | 0 | 0 | 0 |
| **00025465d4725e87** | "  Congratulations from me as well, use the tools well.  · talk " | 0 | 0 | 0 | 0 | 0 | 0 |
| **00031b1e95af7921** | Your vandalism to the Matt Shirvington article has been reverted.  Please don't do it again, or you will be banned. | 0 | 0 | 0 | 0 | 0 | 0 |
| **00037261f536c51d** | Sorry if the word 'nonsense' was offensive to you. Anyway, I'm not intending to write anything in the article(wow they would jump on me for vandalism), I'm merely requesting that it be more encyclopedic so one can use it for school as a reference. I have been to the selective breeding page but it's almost a stub. It points to 'animal breeding' which is a short messy article that gives you no info. There must be someone around with expertise in eugenics? 93.161.107.169 | 0 | 0 | 0 | 0 | 0 | 0 |
| **00040093b2687caa** | alignment on this subject and which are contrary to those of DuLithgow | 0 | 0 | 0 | 0 | 0 | 0 |
| **0005300084f90edc** | "  Fair use rationale for Image:Wonju.jpg  Thanks for uploading Image:Wonju.jpg. I notice the image page specifies that the image is being used under fair use but there is no explanation or rationale as to why its use in Wikipedia articles constitutes fair use. In addition to the boilerplate fair use template, you must also write out on the image description page a specific explanation or rationale for why using this image in each article is consistent with fair use.  Please go to the image description page and edit it to include a fair use rationale.  If you have uploaded other fair use media, consider checking that you have specified the fair use rationale on those pages too. You can find a list of 'image' pages you have edited by clicking on the ""my contributions"" link (it is located at the very top of any Wikipedia page when you are logged in), and then selecting ""Image"" from the dropdown box. Note that any fair use images uploaded after 4 May, 2006, and lacking such an explanation will be deleted one week after they have been uploaded, as described on criteria for speedy deletion. If you have any questions please ask them at the Media copyright questions page. Thank you. (talk • contribs • )  Unspecified source for Image:Wonju.jpg  Thanks for uploading Image:Wonju.jpg. I noticed that the file's description page currently doesn't specify who created the content, so the copyright status is unclear. If you did not create this file yourself, then you will need to specify the owner of the copyright. If you obtained it from a website, then a link to the website from which it was taken, together with a restatement of that website's terms of use of its content, is usually sufficient information. However, if the copyright holder is different from the website's publisher, then their copyright should also be acknowledged.  As well as adding the source, please add a proper copyright licensing tag if the file doesn't have one already. If you created/took the picture, audio, or video then the  tag can be used to release it under the GFDL. If you believe the media meets the criteria at Wikipedia:Fair use, use a tag such as  or one of the other tags listed at Wikipedia:Image copyright tags#Fair use. See Wikipedia:Image copyright tags for the full list of copyright tags that you can use.  If you have uploaded other files, consider checking that you have specified their source and tagged them, too. You can find a list of files you have uploaded by following [ this link]. Unsourced and untagged images may be deleted one week after they have been tagged, as described on criteria for speedy deletion. If the image is copyrighted under a non-free license (per Wikipedia:Fair use) then the image will be deleted 48 hours after . If you have any questions please ask them at the Media copyright questions page. Thank you. (talk • contribs • ) " | 0 | 0 | 0 | 0 | 0 | 0 |
| **00054a5e18b50dd4** | bbq  be a man and lets discuss it-maybe over the phone? | 0 | 0 | 0 | 0 | 0 | 0 |
| **0005c987bdfc9d4b** | Hey... what is it..  @ | talk .  What is it... an exclusive group of some WP TALIBANS...who are good at destroying, self-appointed purist who GANG UP any one who asks them questions abt their ANTI-SOCIAL and DESTRUCTIVE (non)-contribution at WP?  Ask Sityush to clean up his behavior than issue me nonsensical warnings... | 1 | 0 | 0 | 0 | 0 | 0 |
| **0006f16e4e9f292e** | Before you start throwing accusations and warnings at me, lets review the edit itself-making ad hominem attacks isn't going to strengthen your argument, it will merely make it look like you are abusing your power as an admin.  Now, the edit itself is relevant-this is probably the single most talked about event int he news as of late. His absence is notable, since he is the only living ex-president who did not attend. That's certainly more notable than his dedicating an aircracft carrier.  I intend to revert this edit, in hopes of attracting the attention of an admin that is willing to look at the issue itself, and not throw accusations around quite so liberally. Perhaps, if you achieve a level of civility where you can do this, we can have a rational discussion on the topic and resolve the matter peacefully. | 0 | 0 | 0 | 0 | 0 | 0 |
| **00070ef96486d6f9** | Oh, and the girl above started her arguments with me. She stuck her nose where it doesn't belong. I believe the argument was between me and Yvesnimmo. But like I said, the situation was settled and I apologized. Thanks, | 0 | 0 | 0 | 0 | 0 | 0 |
| **00078f8ce7eb276d** | "  Juelz Santanas Age  In 2002, Juelz Santana was 18 years old, then came February 18th, which makes Juelz turn 19 making songs with The Diplomats. The third neff to be signed to Cam's label under Roc A Fella. In 2003, he was 20 years old coming out with his own singles ""Santana's Town"" and ""Down"". So yes, he is born in 1983. He really is, how could he be older then Lloyd Banks? And how could he be 22 when his birthday passed? The homie neff is 23 years old. 1983 - 2006 (Juelz death, god forbid if your thinking about that) equals 23. Go to your caculator and stop changing his year of birth. My god." | 0 | 0 | 0 | 0 | 0 | 0 |
| **000897889268bc93** | REDIRECT Talk:Voydan Pop Georgiev- Chernodrinski | 0 | 0 | 0 | 0 | 0 | 0 |

1. **Tentative list of Algorithms :**

**KERAS LSTM**

The approach that we are taking is to feed the comments into the LSTM as part of the neural network but we can't just feed the words as it is.

So this is what we are going to do:

1. Tokenization - We need to break down the sentence into unique words. For eg, "I love cats and love dogs" will become ["I","love","cats","and","dogs"]
2. Indexing - We put the words in a dictionary-like structure and give them an index each For eg, {1:"I",2:"love",3:"cats",4:"and",5:"dogs"}
3. Index Representation- We could represent the sequence of words in the comments in the form of index, and feed this chain of index into our LSTM. For eg, [1,2,3,4,2,5]
4. Logistic regression with words
5. pretrained embeddings like Word2Vec, GLOVE and Fasttext

**Word2Vec**

1. The main idea behind it is that you train a model on the context on each word, so similar words will have similar numerical representations.
2. Just like a normal feed-forward densely connected neural network(NN) where you have a set of independent variables and a target dependent variable that you are trying to predict, you first break your sentence into words(tokenize) and create a number of pairs of words, depending on the window size. So one of the combination could be a pair of words such as ('cat','purr'), where cat is the independent variable(X) and 'purr' is the target dependent variable(Y) we are aiming to predict.
3. We feed the 'cat' into the NN through an embedding layer initialized with random weights, and pass it through the softmax layer with ultimate aim of predicting 'purr'. The optimization method such as SGD minimize the loss function "(target word | context words)" which seeks to minimize the loss of predicting the target words given the context words. If we do this with enough epochs, the weights in the embedding layer would eventually represent the vocabulary of word vectors, which is the "coordinates" of the words in this geometric vector space.

A screenshot of a cell phone

Description automatically generated

**GLOVE**

GLOVE works similarly as Word2Vec. While you can see above that Word2Vec is a "predictive" model that predicts context given word, GLOVE learns by constructing a co-occurrence matrix (words X context) that basically count how frequently a word appears in a context. Since it's going to be a gigantic matrix, we factorize this matrix to achieve a lower-dimension representation. There's a lot of details that goes in GLOVE but that's the rough idea.

**FastText**

FastText is quite different from the above 2 embeddings. While Word2Vec and GLOVE treats each word as the smallest unit to train on, FastText uses n-gram characters as the smallest unit. For example, the word vector ,"apple", could be broken down into separate word vectors units as "ap","app","ple". The biggest benefit of using FastText is that it generate better word embeddings for rare words, or even words not seen during training because the n-gram character vectors are shared with other words. This is something that Word2Vec and GLOVE cannot achieve.

**5.References :**

1. <https://arxiv.org/pdf/1903.06765.pdf>
2. <https://www.researchgate.net/publication/331857893_A_Machine_Learning_Approach_to_Comment_Toxicity_Classification>
3. <https://arxiv.org/pdf/1810.01869.pdf>
4. <https://www.researchgate.net/publication/327345300_Challenges_for_Toxic_Comment_Classification_An_In-Depth_Error_Analysis>
5. <https://aip.scitation.org/doi/pdf/10.1063/1.5082126>
6. <https://arxiv.org/pdf/1903.06765.pdf>
7. <https://www.researchgate.net/profile/Hetal_Patel27/publication/324169011_Toxic_CTools-A-Case-Study.pdf>
8. <https://www.rsisinternational.org/journals/ijrias/DigitalLibrary/Vol.4&Issue11/142-147.pdf>
9. <https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/reports/6856482.pdf>
10. <http://towardsdatascience.com/attack-toxic-comments-kaggle-competition-using-fast-ai-b9eb61509e79>
11. <https://mc.ai/jigsaw-unintended-bias-toxic-comment-classification/>
12. <https://medium.com/google-design/tuning-out-toxic-comments-with-the-help-of-ai-85d0f92414db>
13. <https://www.perspectiveapi.com/#/home>
14. <https://docs.coralproject.net/talk/toxic-comments/>
15. <https://nycdatascience.com/blog/student-works/toxic-comments-classification-and-non-toxic-chat-application/>
16. <https://link.springer.com/article/10.1186/s13673-019-0205-6>
17. <http://ceur-ws.org/Vol-2353/paper25.pdf>
18. <https://www.technology.org/2020/04/30/jigsaw-multilingual-toxic-comment-classification/>
19. <https://conversationai.github.io/research.html>
20. <https://thenextweb.com/artificial-intelligence/2018/09/11/googles-hate-speech-ai-easily-fooled/>
21. <https://venturebeat.com/2020/01/13/carnegie-mellon-researchers-propose-ai-that-surfaces-positive-online-comments/>
22. <https://conversationai.github.io/research.html>
23. <https://venturebeat.com/2020/01/13/carnegie-mellon-researchers-propose-ai-that-surfaces-positive-online-comments/>
24. <https://en.wikipedia.org/wiki/Artificial_intelligence_in_Wikimedia_projects>
25. <http://aclweb.org/anthology/W18-4412.pdf>
26. <https://thenextweb.com/artificial-intelligence/2018/09/11/googles-hate-speech-ai-easily-fooled/>
27. <http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26647691.pdf>
28. <https://www.theregister.co.uk/2018/08/31/ai_toxic_comments/>
29. <https://ai.googleblog.com/2019/12/fairness-indicators-scalable.html>
30. <https://opendatascience.com/data-science-news-week-in-review-november-25th/>
31. <https://www.forbes.com/sites/bernardmarr/2018/08/17/the-amazing-ways-how-wikipedia-uses-artificial-intelligence/#7919f76f2b9d>
32. <http://medium.com/the-false-positive/increasing-transparency-in-machine-learning-models-311ee08ca58a>