NSFW Images & Profane Text Filter

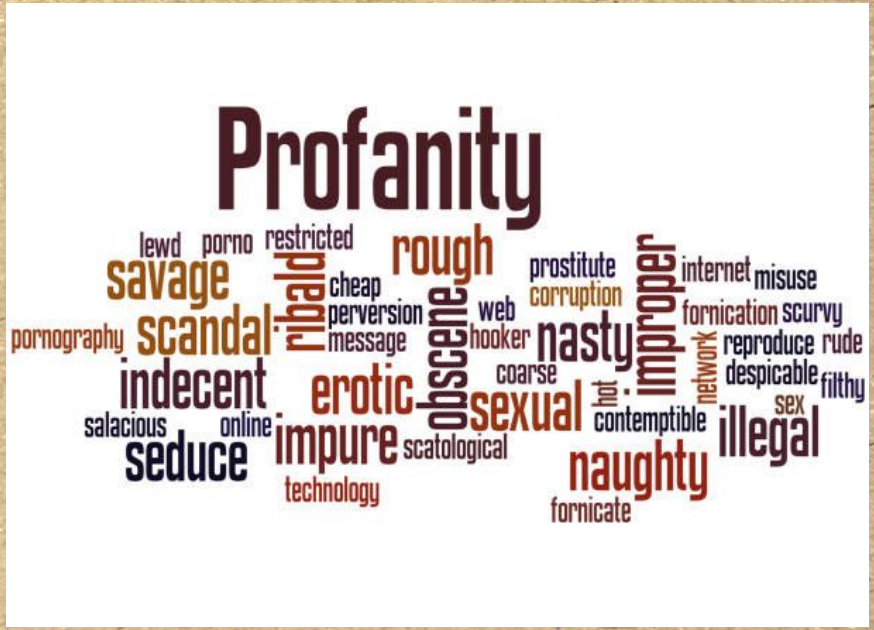
Team 1
Anchit Gupta (20171041)
Neel Trivedi (20171015)
Kunal Vaswani (20171068)

Links
Github Repo
Demo Video

The “Social” Part

What is NSFW and Profane content

- There are lots of different category of content currently on internet which can be classified into NSFW (not safe for work) and profane.
- This content exists in all forms of media be it textual, images, video or audio.
- Some of the classes of such content include,
  - Pornographic/Sexual images
  - Erotic content
  - Violent or violence inducing images and video
  - Animal Cruelty related content
  - Hateful and abusive texts
  - Sexist or racist content
  - ...and much more



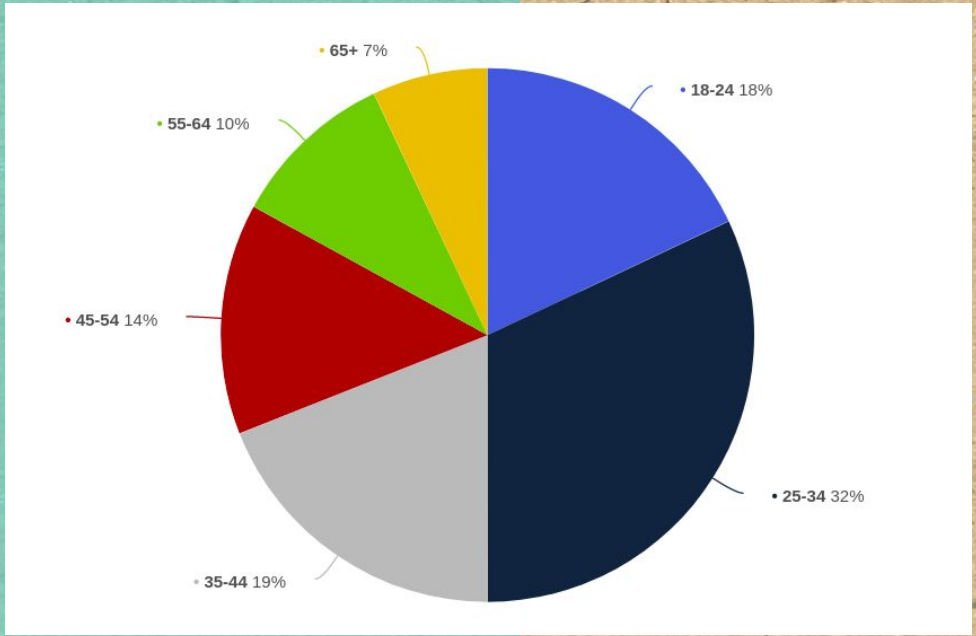

Tweets
Tweets & replies
Media
KAPIL @KapilSharmaK9 · 10m
Maa chuda Gaya yahan ka system.. Saale ghatiya log .. Agar main prime minister hota to fake news banane walo ko faansi laga deta.. saale ghatiya
707 424 969
KAPIL @KapilSharmaK9 · 16m
According to sources this is the news .. u motherfucker why don't u tell who r ur sources
371 330 1.0K
KAPIL @KapilSharmaK9 · 18m
N a request to media.. pls don't make it negative news just to sell ur paper ... he is a nice man n he will come@ out of it soon.इतने बड़े बड़े घोटाले हो हुए.. तब तो तुम बोले नहीं..कितना लेते हो negative न्यूज़ स्प्रेड करने के liye.u fucking paid media ..specially @Spotboye u mc
271 395 1.1K
KAPIL @KapilSharmaK9 · 23m
मैंने बहुत सारे ऐसे महाराजा टाइप लोग देखे हैं जो बड़े फ़ख्र से बताते हैं की हमने शेरकाशिकार किया .. मैं मिला हु उनसे. सलमान बहुत लोगों की मदद करता है.. अच्छा आदमी है..I don't know if he did it or not .. but see his best sides.. ghatiya system .. let me do good work ..


Jul 14
Replying to @HackneyAbbott
You forgot "fat disgusting obese chicken-loving nigger"
Jul 14
Replying to @HackneyAbbott
An acid attack would probably make your face look better you fat nigger


Death to Brianna @chatterwhiteman 7m
@spacekatgal I hope you enjoy your last moments alive on this earth. You did nothing worthwhile with your life.
Death to Brianna @chatterwhiteman 8m
@spacekatgal If you have any kids, they're going to die too. I don't give a [redacted]. They'll grow up to be feminists anyway.
Death to Brianna @chatterwhiteman 8m
@spacekatgal Your mutilated corpse will be on the front page of Jezebel tomorrow and there isn't jack [redacted] you can do about it.


Celia @_celia_bedelia_ · 20h
i'm so tired of men.
5 52
William Wallace
@WallWill2000
Replying to @_celia_bedelia_ @Free_Fries_ and 2 others
Go live in a place with no men... oh yeah there is none, then maybe just shut up or shoot yourself.
5/16/18, 9:00 AM

How does such content affect users?



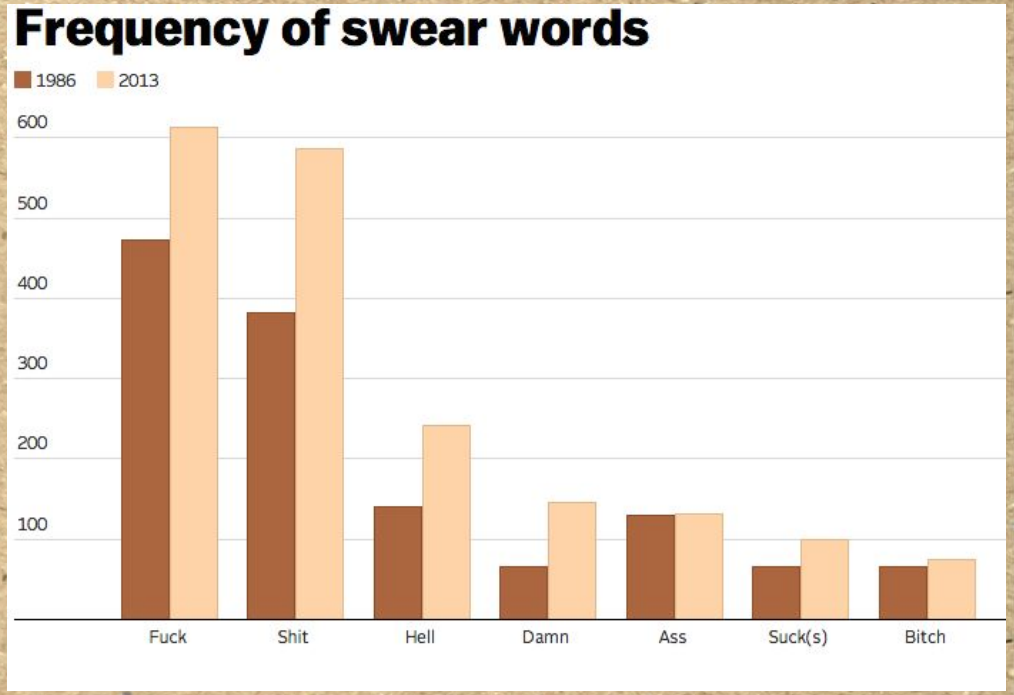
- As can be seen in the plot above, a huge group of internet users are in the age group of 18 to 24. And they are majorly affected by the profanity and NSFW content.

How does such content affect users?

- As we have seen in this course, internet and social media content has a significant impact on the thinking process of the users and especially in the case of young users.
- It has been reported in many parts of the world how terrorist organizations have used violent content like videos of killings and masacre to lure innocent individuals to join their group.
- Online hate and profanity has been known to cause serious mental health issues among teenagers as reported by many studies.
- Depression and anxiety caused by mass hate on social media platform has led to cases on suicide as well.
- As shown by many research studies, constant engagement to such content has led students to be highly distracted from studies.

Why should such content be filtered?

- The plot here clearly shows how profanity and hate text is constant increasing on the internet.
- As mentioned in the previous slides, such profane and NSFW content has significant negative effect on users.
- With internet and social media becoming omnipresent in today's world, it seems necessary that the content that they serve is checked.
- Many popup ads and clickbaits also contain such unwanted content and leads to distraction and inconvenience to the users and hence filtering such content can lead to better online experience.



The “Computing” Part

Real-time solution

Our aim here is to build a browser extension and thus obtaining solutions in real-time is crucial here.

We thus limit our testing to cpus and use the following hardware:

Processor: Intel(R) Core(TM) i7-7700HQ CPU
Installed RAM: 8.00 GB (7.89 GB usable)
System type: 64-bit operating system, x64-based processor

NSFW Image Classification

Problem Definition

Not-Safe-For-Work images can be described as any images which can be deemed inappropriate in a workplace primarily because it may contain:

- Sexual or pornographic images
- Violence
- Extreme graphics like gore or abusive
- Suggestive content

Platforms such as YouTube/Instagram/LinkedIn etc., where anyone can upload images or videos, it is hard to keep the platforms safe, *especially from kids*. Millions of images are uploaded to these platforms every day, and verifying every image manually is an impossible task. Therefore we use deep learning to solve this problem.

Dataset Collection

Unfortunately no other data apart from sexual/pornographic images were available for this task.

We have collected the data from multiple sources:

- https://www.kaggle.com/omeret/not-safe-for-work/version/1
- https://github.com/EBazarov/nsfw_data_source_urls

The dataset had to be cleaned after downloading, for example:

- Delete duplicate images
- Delete corrupt images
- Delete images with 0 size

Architecture

We tried various CNN based deep learning models for NSFW image classification.

- Transfer learning + fully connected layer
  - ResNet50
  - ResNext101
  - VGG
  - MobileNetV2

- Model trained from Scratch
  - 5 cnn layers + 3 fully connected layers

Results

Model	Accuracy	Inference time for 1 image
MobileNetV2	96.35	52ms
ResNet50	91.28	82ms
ResNeXt101	92	180ms
VGG	90.78	106ms
Ours from scratch	71.69	30ms

We decided to use MobileNetV2 for our browser extension as it has highest accuracy and also lesser inference time.

Profane Text Detection

Problem Definition

Detect profanity in text-strings.

Sentence	Is it profane?
Hello, how are you?	No
Fuck you	Yes

Dataset Collection

1. https://github.com/t-davidson/hate-speech-and-offensive-language/tree/master/data.
2. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

These dataset were based on more complicated studies like hate speech detection.
Deeper classes in dataset: toxicity, threat, hate speech, offensive language.

Profane: Data with any kind of offensive language.
Non-profane: Rest.

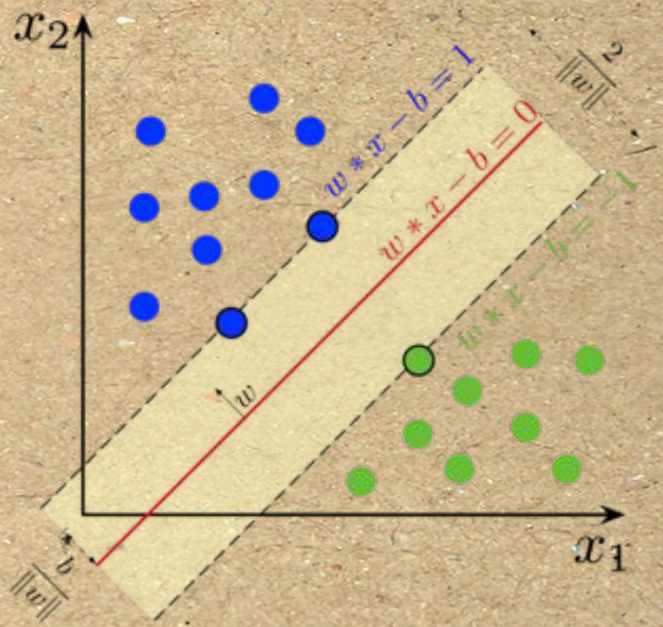
Fastest Approach!

We start with simplest strategies for our first approach.

Vectorization of sentences: CountVectorizer class

Model: scikit-learn LinearSVC class (SVM model)

Reason: Fast enough to run in real-time yet robust enough to handle many different kinds of profanity.



Results

Testing Accuracy: 95.28

Really FAST!

No. of Samples	Time taken to predict
500	0.30s
1500	0.88s
3500	2.18s

Issues

Contextual information is missing in Vectorization of Sentences!

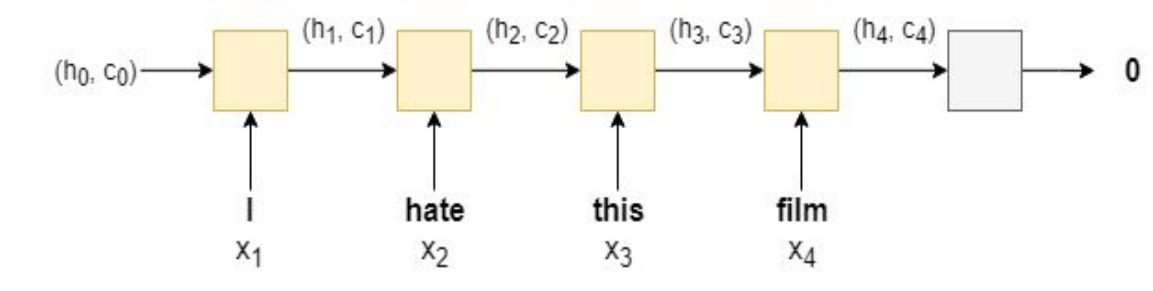
Sentence 1 : “I hate how much I love this”

Sentence 2: “I love how much I hate this”

Sentence 1 and 2 has similar word counts and one can’t differentiate them without sequential information.
Eg: “Many aphides, &c., puncture the leaves, suck out the sap, and induce va:ious local deformations, arrest..”

We are training our model end-to-end and thus missing out on transfer learning.

LSTMs



To incorporate sequential information we use the lstm model.

We resist ourselves from using simple RNNs since they suffer from vanishing gradient problems.

We also exploit further tricks by making our lstm model bidirectional and multi layer.

Pre-trained Word Embeddings

The main intuition on why pre-trained embeddings help is that words with similar semantic meaning would be close together in vector space.

We use pre trained 100 dimension glove embeddings trained on 6 billion tokens.

GloVe: Global Vectors for Word Representation

Jeffrey Pennington, Richard Socher, Christopher D. Manning
Computer Science Department, Stanford University, Stanford, CA 94305
jpennin@stanford.edu, richard@socher.org, manning@stanford.edu

Results

Training Accuracy	96.83
Testing Accuracy	96.17

Time evaluations:

No. of Samples	Time taken to predict
50	1.76s
100	3.81s
300	12.18s

Misssssspellings!

LSTMS rely on a vocabulary space and thus face issues with misspellings!

Misspellings are out of vocabulary words! Words like “f**k”, “biiiiiiitch” are difficult to approach without using subword information.

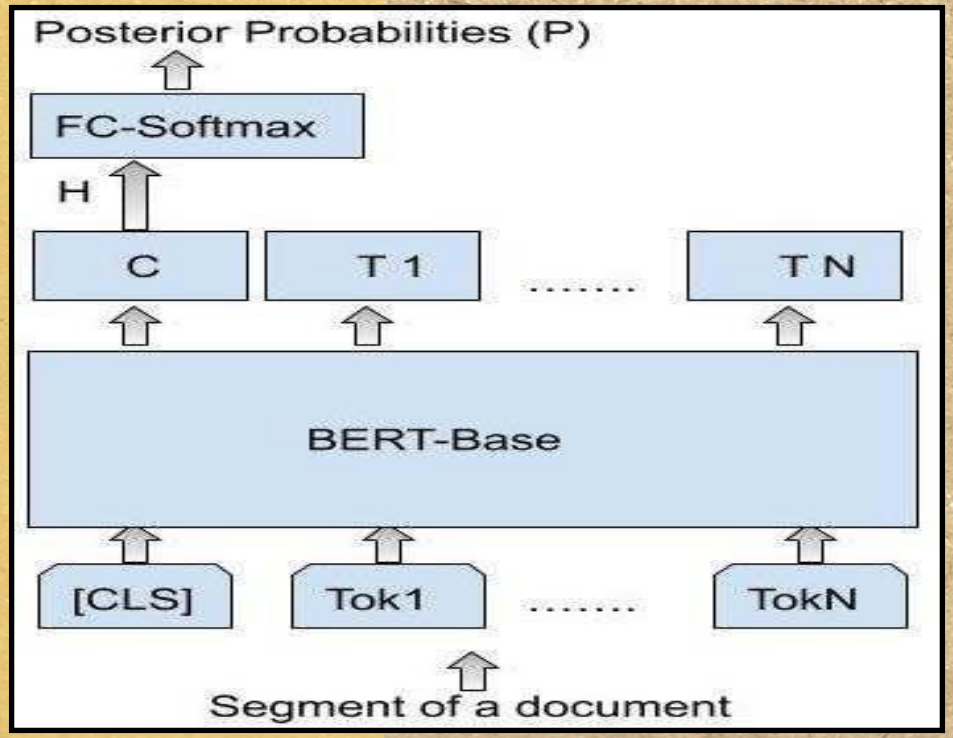
Approach: use BERT! :)

BERT exploits subword information using WordPiece. WordPiece is a subword segmentation algorithm used in natural language processing.

Other reasons for superior performance for BERT:

- Masked language modelling
- At its heart it uses transformers! Lot of ATTENTION

Model


Posterior Probabilities (P)
FC-Softmax
H
C
T 1
.....
T N
BERT-Base
[CLS]
Tok1
.....
TokN
Segment of a document

Results

Can be increased a lot given a lot of training time.

We also trained it on 1 gpu and thus didn't use much parallelization.

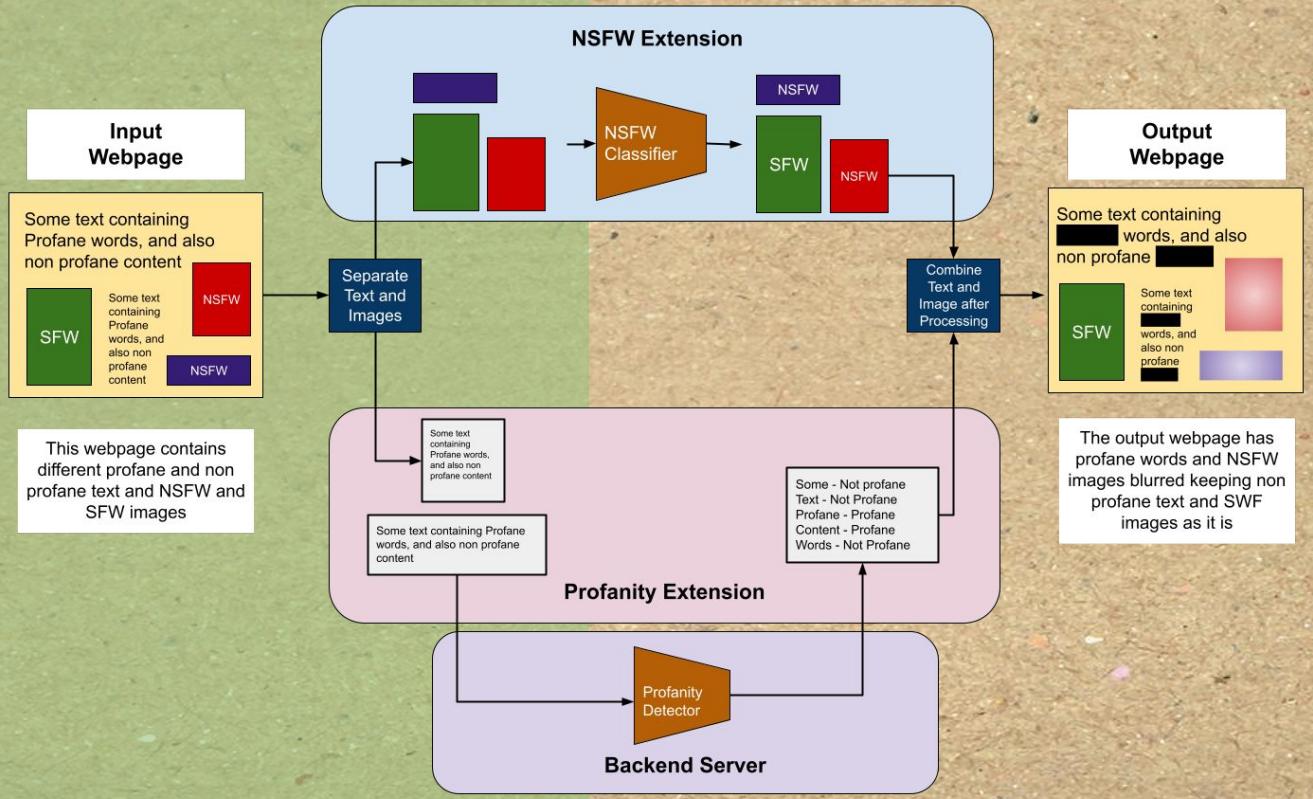
Training Accuracy	94.55
Testing Accuracy	94.05

Time evaluations:

No. of Samples	Time taken to predict
50	7.38s
100	15.98s
300	42.68s

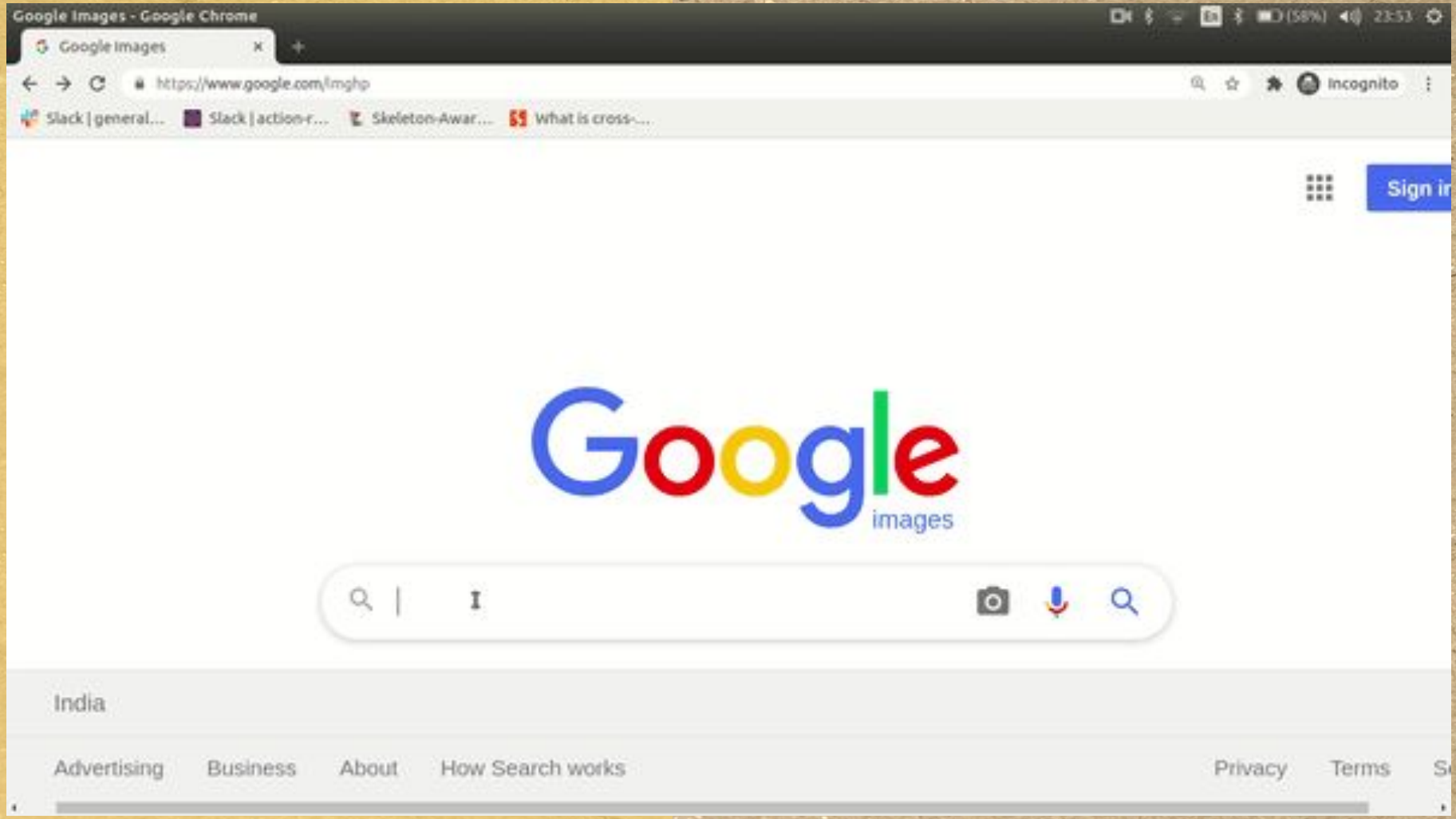
Browser Extension

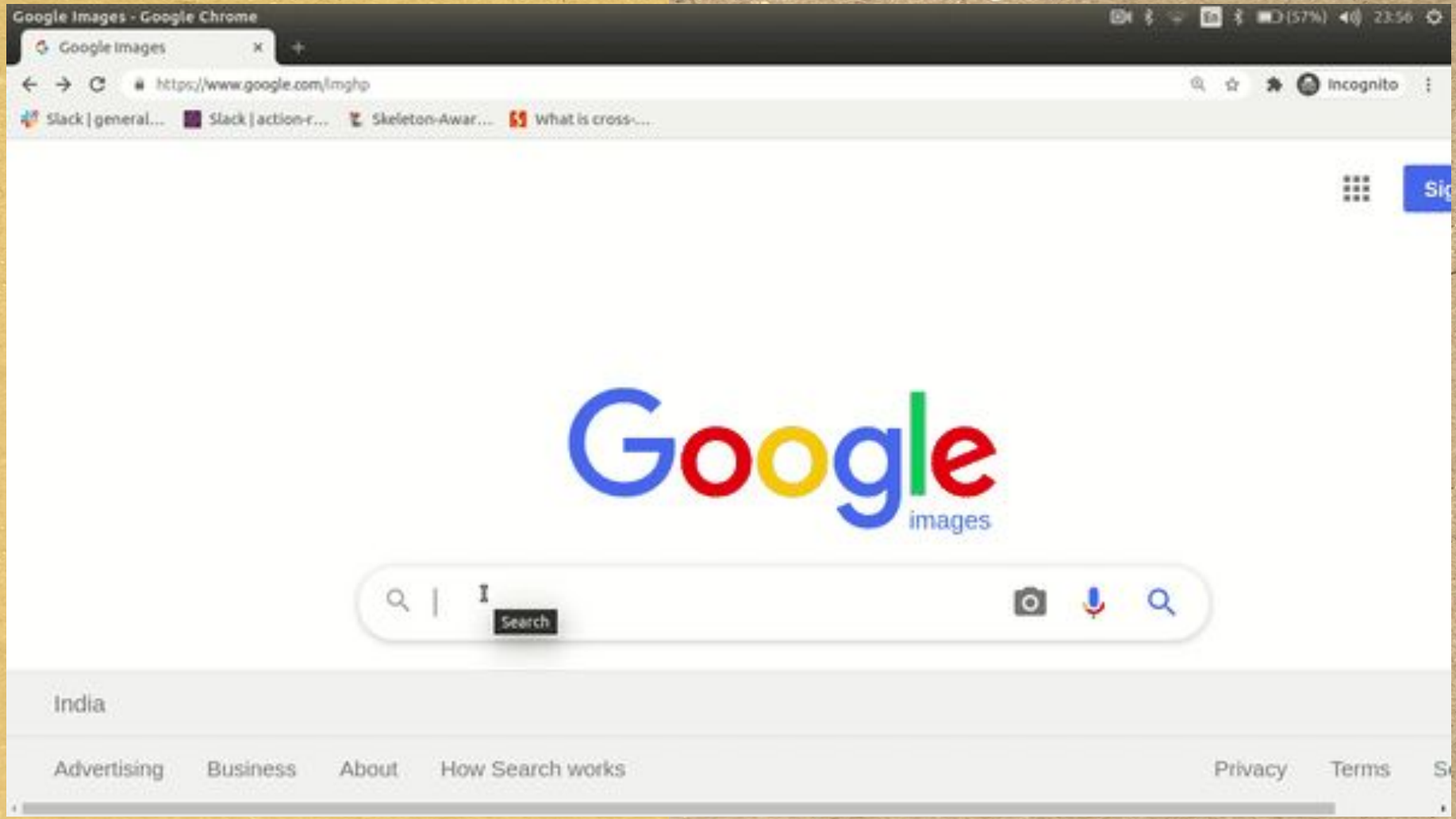
Architecture Diagram



Extension Mechanism

- As shown in the use case diagram in the previous slide, the browser extension captures all the images from the webpage and puts them into a queue.
- The pretrained model is used to classify the image into NSFW or SFW and those probabilities are returned.
- The images which are classified as NSFW are blurred using a filter for that image component.
- Similar process will be used for the text of webpage, where we will be adding text components into the queue and use pretrained models to determine whether the text is profane or not.
- However processing text on browser directly wasn't possible due to speed and chrome restrictions. So we have created a backend server which analyzes text and updates the webpage


Google Images - Google Chrome
Google Images
https://www.google.com/imghp
Incognito
Slack | general...
Slack | action-r...
Skeleton-Awar...
What is cross-...
Sign in
Google
images
India
Advertising
Business
About
How Search works
Privacy
Terms


Google Images - Google Chrome
Google Images
https://www.google.com/imghp
Incognito
Slack | general...
Slack | action-r...
Skeleton-Awar...
What is cross-...
Google
images
Search
India
Advertising
Business
About
How Search works
Privacy
Terms

Without Extension

This is some non - profane text

This extension is created by Neel, Anchit and Kunal.

This is profane text

This is fucking shit. I will kill you bitch

Some more profane text example

Fuck you bitch

You cunt

With Extension

This is some non - profane text

This extension is created by Neel, Anchit and Kunal.

This is profane text

T**********

Some more profane text example

F**********

Y*******

Project Timeline

28th Jan

Did literature survey on existing methods and available datasets for the task of NSFW image classification and profanity detection

15th Feb

Defined problem statement, expected outcome and solution approach

15th Mar

Implemented first set of classifier and text detectors for NSFW classification and profanity detection

1st April

Implemented working browser extension for NSFW image classification

17th April

Add profanity detector to the browser extension and make final submission

Thank You