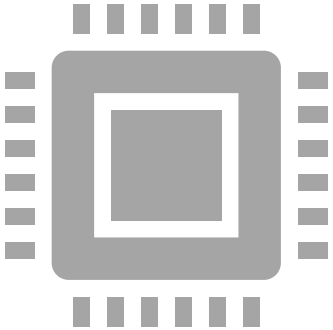
IMAGE SPAM BUSTER

Team Name

Sudo Alliance

What is an Image based spam ?

Trick that consists of embedding the spam message into attached images, to prevent its detection by text-based filters.

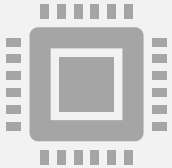


Sometimes the text embedded into images is obfuscated, to undermine OCR-based filters.

Image Spams at Mall91

Why SPAM Buster is necessary?

At the current state of the world, thousand of million people are connected together through electronic devices and they are involved in high volume and various collaborations around the world. Utilizing the same platform, some peoples are making this to use for their promotional job as well as for spreading nudity.



To make our platform safe, we are using SPAM Buster based on cutting edge techniques to make our platform safe.

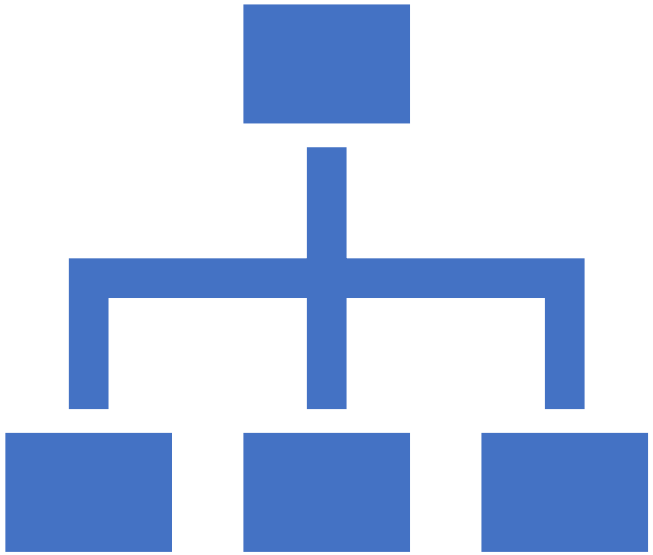
Spam Buster? How ?

Image Spams Type

Image based Spams can be in two forms:

Spam purely based on Image

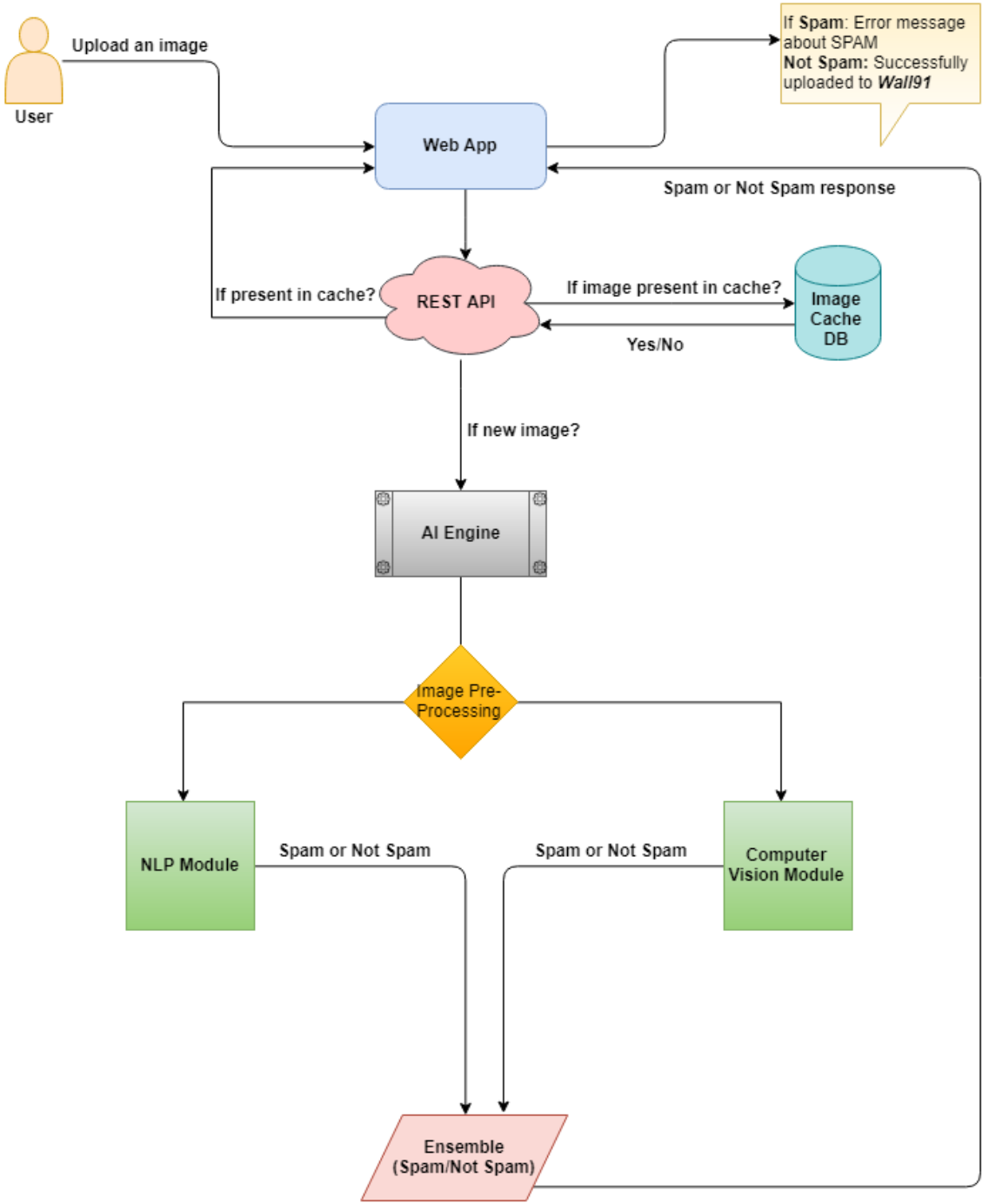
Spam based on Text present in the image



Need of a Multi-modal approach

- There needs to be a system which leverages both on textual features and Image based features to successfully bust the Spams.
- Both Mode of approaches can complement each other

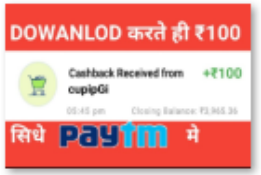
Design Overview



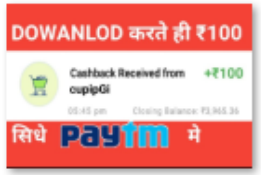
Data Challenges – Duplicate Images

15

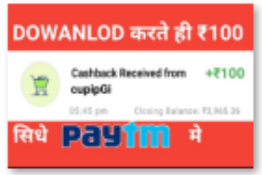
16



17



18



19

25

26

27

28

29

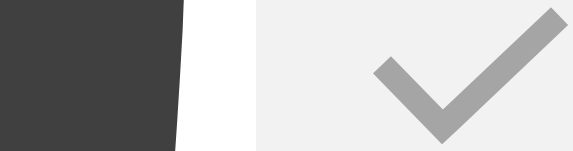
Data Challenges – Mislabeled Images

Labelled as Spam

Labelled as a genuine post

Overcoming Data Challenges

Image De-Duplication Algorithm



Removing Mislabeled images at scale

Approaches

NLP Approach

- OCR
- Heuristics
- Text classifier

Computer Vision Approach

- Baseline classifier (SVM)
- Resnet 50 based CNN

Text Based Approach

1

Text content based spam is one of the most prominent types of Spam nowadays.

2

We need to first process the image and extract all possible texts through OCR engine.

3

The extracted text is then used in heuristic model as well as logistic regression.

OCR

OCR OR OPTICAL CHARACTER RECOGNITION IS THE PROCESS OF EXTRACTION OF TEXT FROM A GIVEN IMAGE.

BEFORE EXTRACTING TEXT FROM IMAGE WE PRE-PROCESS DATA BY SCALING, CROPPING AND DE-SKEWING.

ONCE WE HAVE THE FORMATTED IMAGE WE ARE USING PYTESSERACT, WHICH IS A WRAPPER OF GOOGLE TESSERACT-OCR LIBRARY, FOR TEXT EXTRACTION.

Heuristic Approach

- Heuristic Models refers to techniques based on experience for various tasks such as research, problem solving, discovery and learning
- Given a set of “known-bads” we are grouping them and assigning weights to each group. Based on the weight we are calculating the total score, which if is above a set threshold will be classified as a spam.
- The weights and threshold are saved in a json file which is configurable.

Text Classifier

- We generated word vectors of extracted text (English + Hindi) using Tf-IDF approach
- Also, we engineered some features like “total characters in text”, “total tokens in text”, “number of title case tokens”, etc.
- Then a binary classification model is trained using Logistic Regression with Regularization
- We achieved 94% f1-score with this model

Image Based Approach

- Image spams can also be classified without extracting the text present in it.
- We used pre-trained Resnet-50 model along with the provided data to train a binary classifier for Spam vs Non-Spam.
- We used some techniques like Cyclic Learning Rates, Test time Augmentation straight from the latest research papers
- We achieved an f1-score of 91% for this Computer Vision model

Ensemble

- Final approach is to combine the outputs of all the models to come up with a robust Spam Buster
- For now we thought to build this system strict, so if any classifier suggests that image is an spam, ensemble outputs is as an Spam
- We take the mean of all the classifier’s scores to come up with final ensemble score

Web App

Upload Document
History
Analytics
Setting

Select AI Learning Model: Image Classifier

Click or drag file to this area to upload

3.jpg

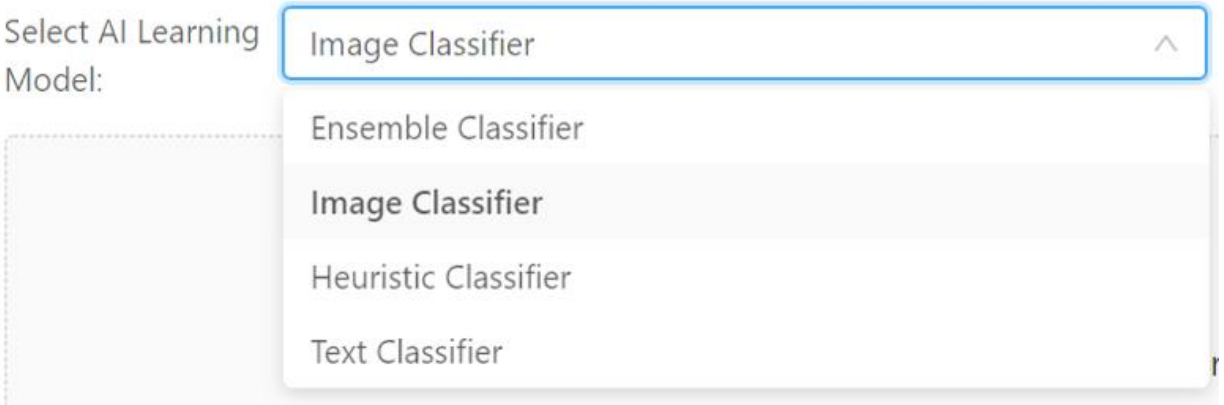
Successfully uploaded

Size: 96.83 KB

Spam: YES

Score: 0.9932170510292053

Web App



- We can choose from any of the four models to get output for an image

Future Scopes

- With more training data available, following approaches can further boost the performance:
  - Train our own word embeddings, then use them to classify the text
  - Use Image Localization followed by OCR for better text extraction
  - Use Attention based CNNs for better image classification

Thanks !

Team Sudo Alliance

GitHub Repo : https://github.com/shwetkm/kreate-hackathon-mall91/

App Link : http://139.59.59.9/spambuster/#/