



GRID 3.0

Team Name

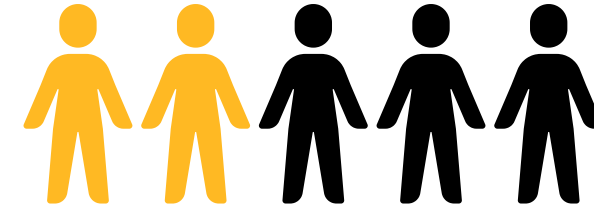
INCOGNITO

(SSASIT, GTU)

The Problem

Less than

10%



Indians have a credit bureau presence, so it is essential to identify and utilize alternate forms of data to assess customers credit worthiness.

To solve this create an explainable trust and affluence score/classification based on all the signals.



Trust is an everyday problem.

It's something that we face all the time.

Lenders are vulnerable to the borrowers, depending on the size of the loan, and they're uncertain about whether the borrowers will be willing and able to repay the loan in the future.

And the thing about credit is that modern economies depend on credit.

Using social media for credit scoring we don't depend so much on face-to-face interactions and direct contact.

Problem Breakdown



Profile DATA Evaluation

Generating the score based on the users profile having different 12+ entities like Bio, Account Creation date, No. of connections etc.



User's Text-based Post Evaluation

Applying ML model to generate trust Score based on all the textual content of the user's profile on social media



Visual Content Analysis

All the visual content like Images are analysed to predict the users trust worthiness score

Use-cases



1

Financial institutions

Financial institutions uses credit scoring to evaluate potential loan default risks. Social media data can be used to strengthen financial institutions predictive power.

2

Event and causes

Social media provide a platform through which users can freely share information. misinformation can play a significant role in the success or failure of an event or a cause.

3

Prediction of user trust

By comprehensive analysis of social media data of users, we can keep track of user's latest information which can be used for user trustworthiness prediction of user.

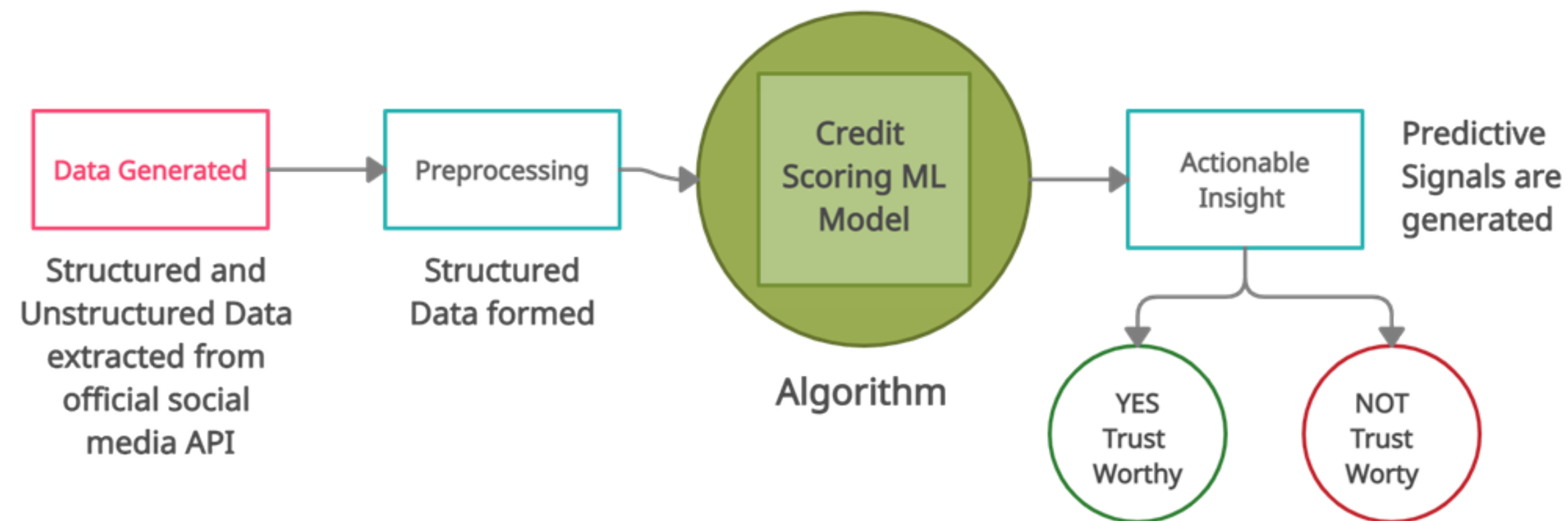
4

Impact on social media

Posts, contents, tweets, reviews, news are uploaded by any trustworthy people or it is just for spreading humor will impact the future of social media the most.

Our Solution

- There will be one system that consists of a Machine Learning Algorithms.
- This system will take input as information of the user for whom we want trustworthiness score and will give output as the user is trustworthy or not.
- User data will pass through all the described stages and at last predicate values will be emitted.



Data Generation

Trusted and official sources



Twitter Developer API



Facebook Developer API

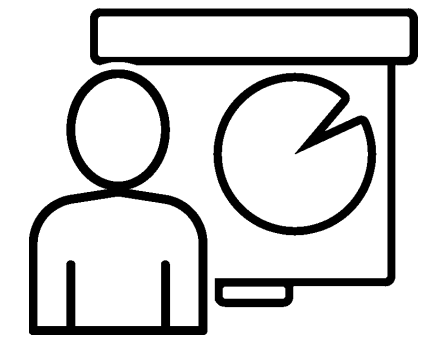
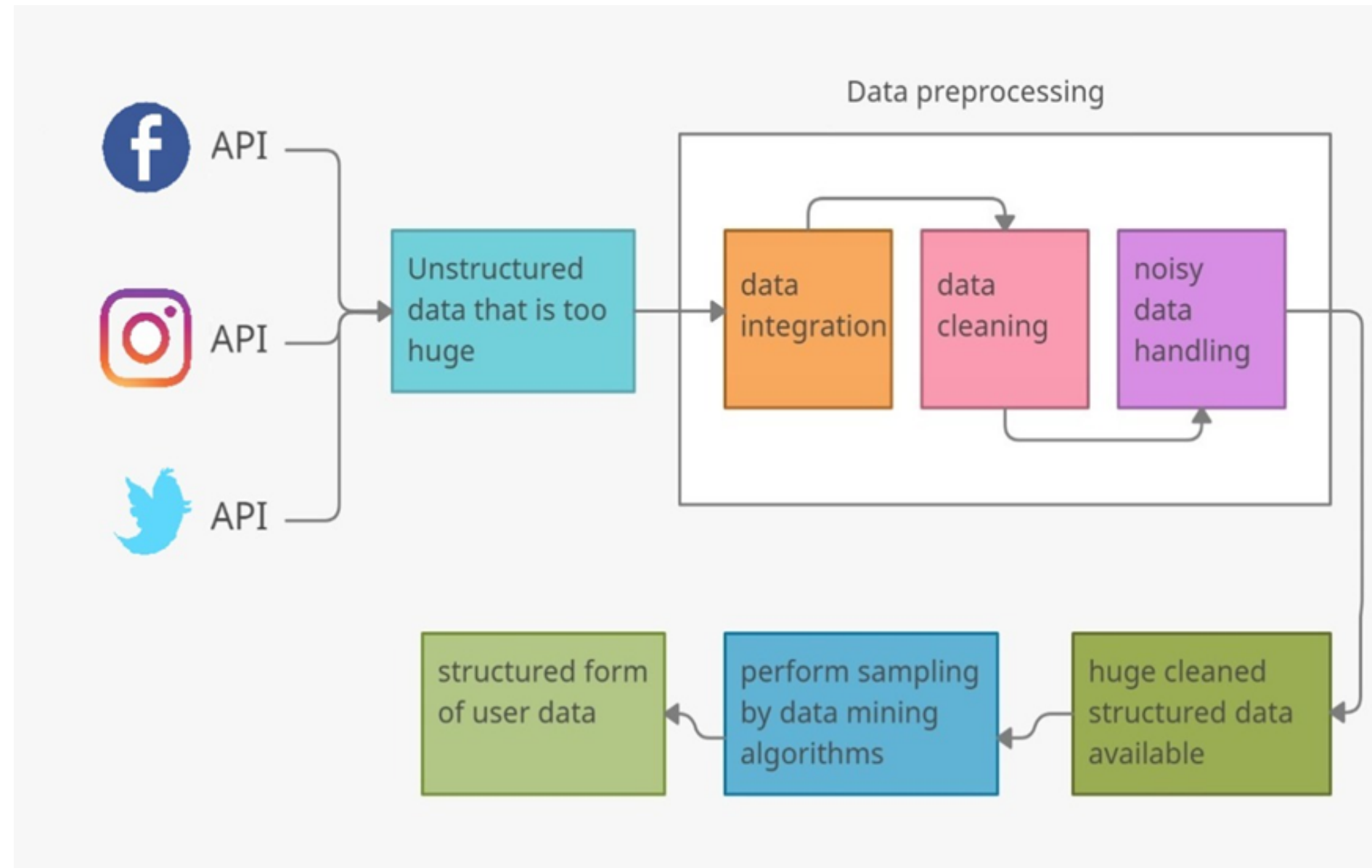
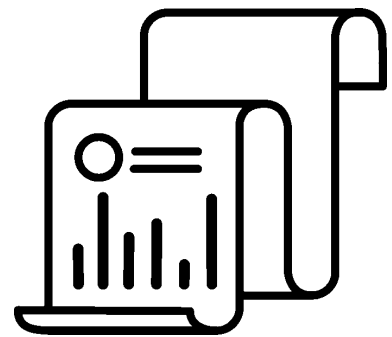


Dataset from trusted
open-sources

- Data will be generated from various official social media API's through API extraction process.
- This data will be in the structured as well as unstructured format on which we will apply to pre-process to make all data aligned and in properly structured format.
- We might get user data from different sources which are either official social Platform API's or the trusted source of labelled dataset.
- Data integration methods require the integration of data from several sources into a single form.

Data Pre-processing

- Generated / Collected data will be noisy and may have an unstructured format means it can have images, speeches, videos, etc that may not be feasible to process using text processing formulas.



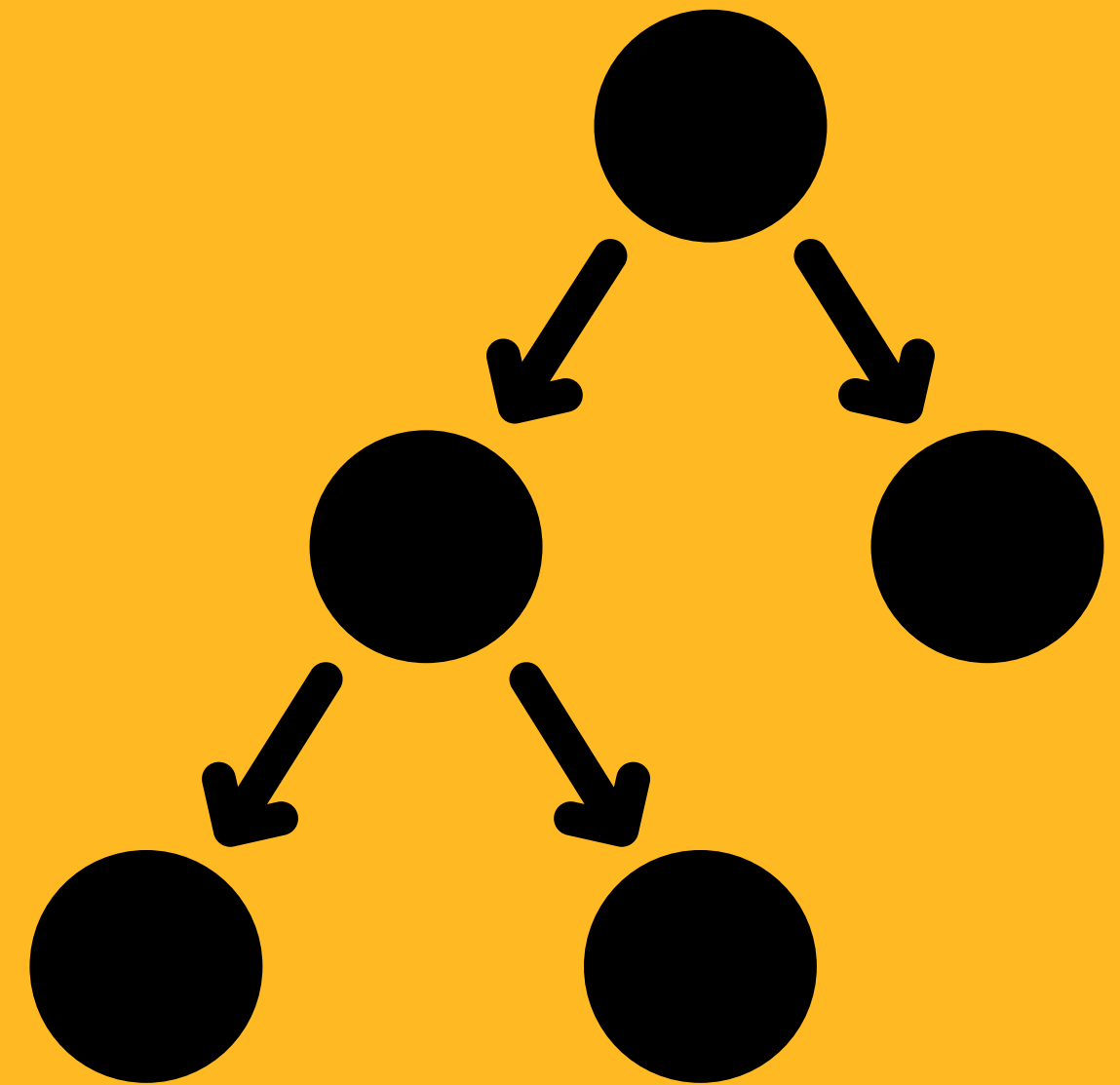


The Algorithm



We studied different supervised machine learning algorithms and came to the conclusion that the

Random forest

ML Algorithm can be used to train our model using **K-fold cross-validation** techniques for the best random train test split and **hyperparameter tuning** is used to find the best parameters and attributes for our model to run it on high accuracy.



Why did we choose Random forest?

1 No overfitting:

Instead of using single decision trees, it will use multiple decision trees and then will give answers as signals from the majority of trees, by this, we have reduced the risk of overfitting.

3 Lesser overall Training Time.

we can run model within couple of seconds and still get the best accuracy

2 Estimates missing data

Random forest is extremely useful when large proportion of data is missing which might be our case

4 High accuracy

With large dataset it produces high accuracy results and even better with parameter tuning

Text-Classification Algorithm

Text Blob and word cloud visualization:

Firstly fetched textual post content and then cleaned that content, after that used text blob python library for sentiment analysis on that content.

The score is decided based on **subjectivity** and **polarity** measures.

Subjectivity

Subjectivity quantifies the amount of personal opinion and factual information contained in the text.

Polarity

It simply means emotions
expressed in a sentence.
which can be positive
negative or neutral.

Visualization of textual content

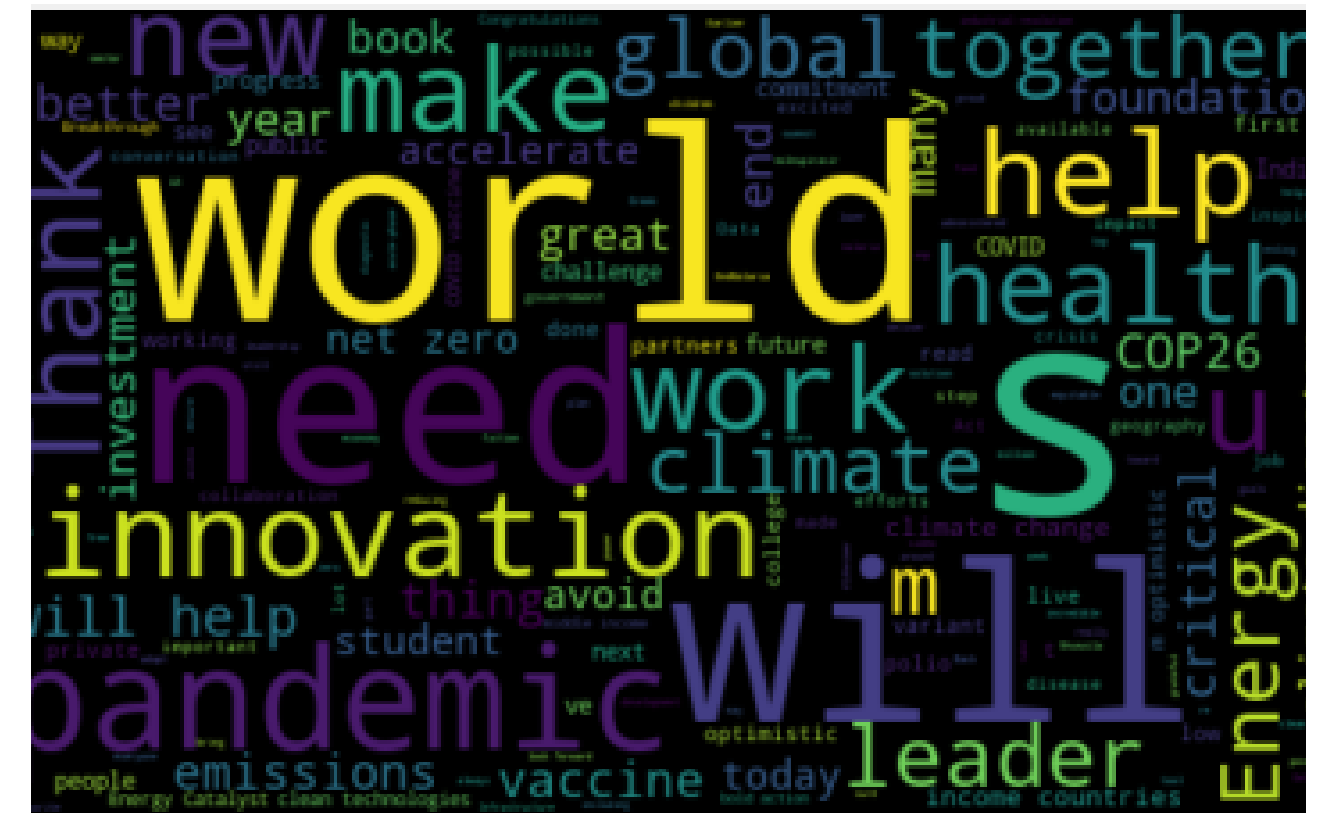
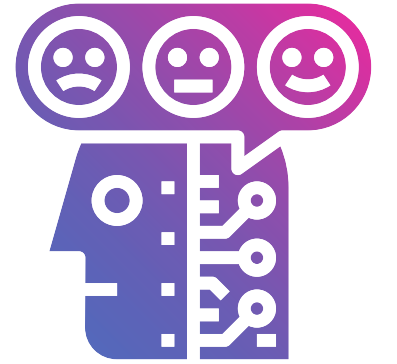


Image Sentiment Analysis



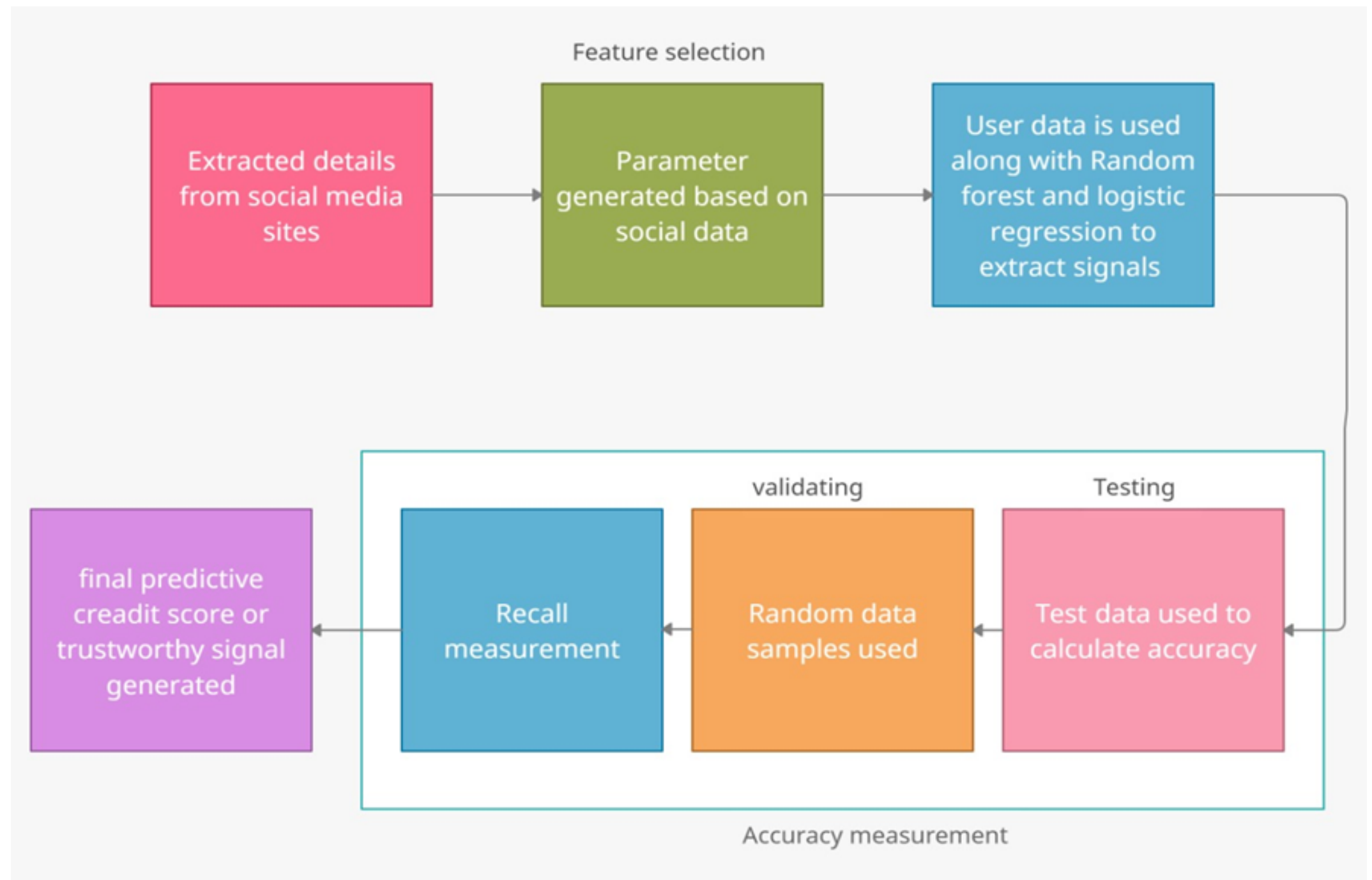
Convolutional Neural Networks and Deep Learning:

- Firstly fetched visual user posts from platforms and then the image sentiment model is built using Tensorflow and Keras Python libraries.
- 2 D images are added layer by layer to a 2D convolutional neural network using a sequential model of Keras.
- for Building multiple layers and matrices of image weights, it needs extensive GPU Processing Power, After training the model we have saved the fitted model into a file using pickle for future running on user data
- In the end, will get precision and recall scores in form of confusion matrices.
- Answer will be = ((average score of all posts) / no of posts) / normalization factor**
here normalization factor is 2 in our case.

How algorithm will work

Most current available credit scoring algorithms use only accuracy scores to evaluate the efficiency of the ML model.

But for our model to increase the efficiency and performance we have used testing (accuracy-score), validating and Recall, F1 scores just to ensure better performance.



Some highlighted features

After extensive studies of different research papers and credit scoring models, we come to the conclusion that the below-mentioned list of features/matrices/indexes can be used to predict the trustworthiness of user.

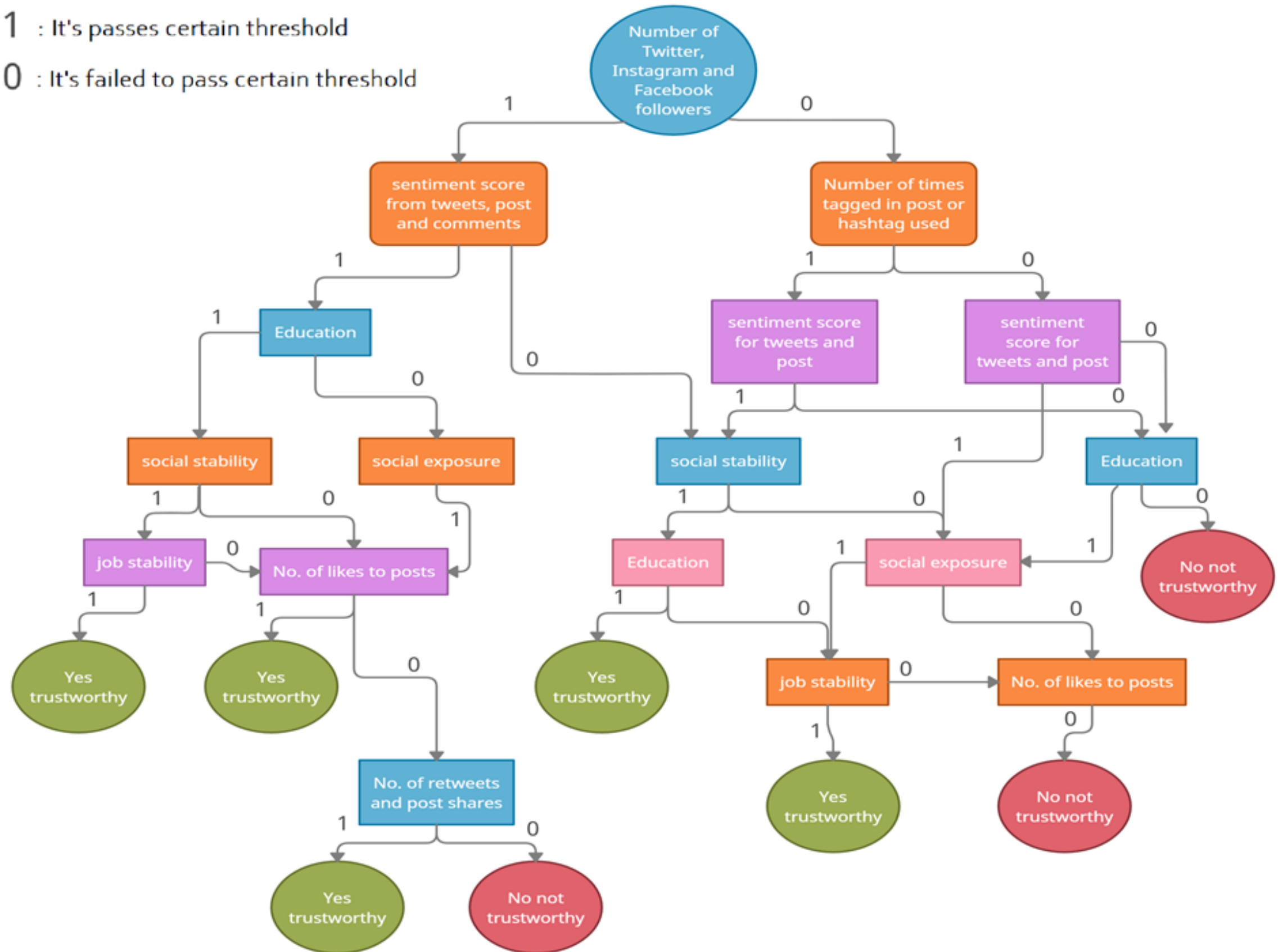
SENTIMENT SCORE	NUMBER OF TIMES TAGGED	NUMBER OF LIKES COUNT	NO. OF FACEBOOK FRIENDS	NO. OF TWITTER FOLLOWERS
SOCIAL STABILITY	PERSONAL BIO	SOCIAL EXPOSURE	NO. OF STATUS UPDATES	DATE OF BIRTH
NUMBER OF MENTIONS	NO. OF FOLLOWINGS	DEFAULT PROFILE PICTURE	NO. OF TIMES POST SHARED	SOCIAL NETWORK QUALITY

Sentiment score will be counted by performing natural language processing NLTK on tweets and comments contents and finding their positive and negative meanings.

A glimpse of random forest

Behind the scenes, several decision trees will be formed by trying different feature combinations and the final score will be evaluated from the overall results.

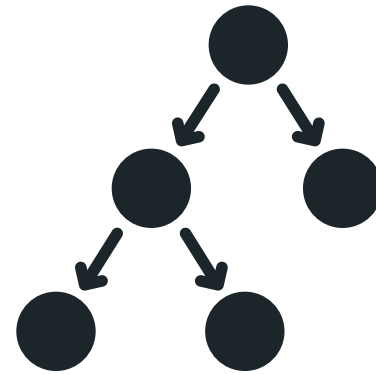
1 : It's passes certain threshold
0 : It's failed to pass certain threshold



Final Trustworthiness score creation

Random Forest
Algorithm

Accuracy: 88.03 %



Text-Classification
Algorithm

Accuracy: 94.65 %

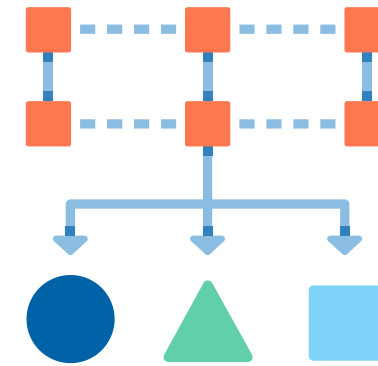
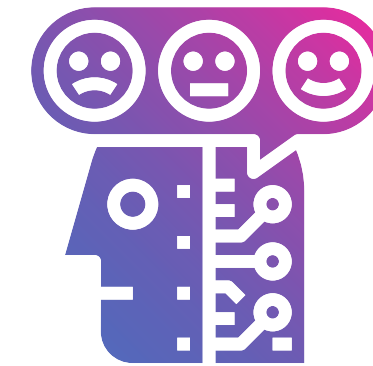


Image Sentiment Analysis
Algorithm

Accuracy: 83.91%



Average



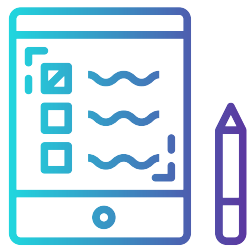
Overall
Accuracy
88.86 % ~ 89%

There is no way by which we can

It may take some time for the algorithm

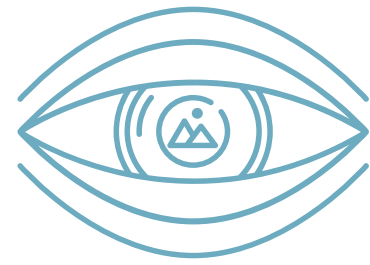
When if any user has entered less

How our system differs from other existing similar systems



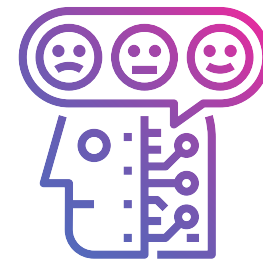
Extended Evaluation

Instead of training and predicting trustworthiness only basis on user profile, we also added one more prediction technique that is user's posts, bio, and tweets classification.



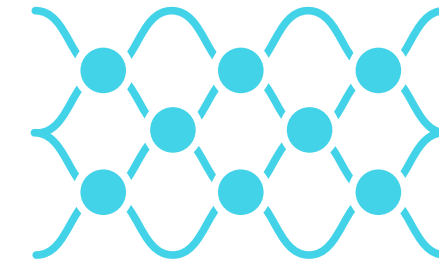
Visual Content

Images are the easiest medium through which people can express their emotions on social networking sites. Social media users are increasingly using images and videos to express their opinions and share their experiences.



Sentiment analysis

Sentiment analysis of such large-scale visual content can help better extract user sentiments toward events or topics, such as those in image tweets so that prediction of sentiment from visual content is complementary to textual sentiment analysis.



Neural networks

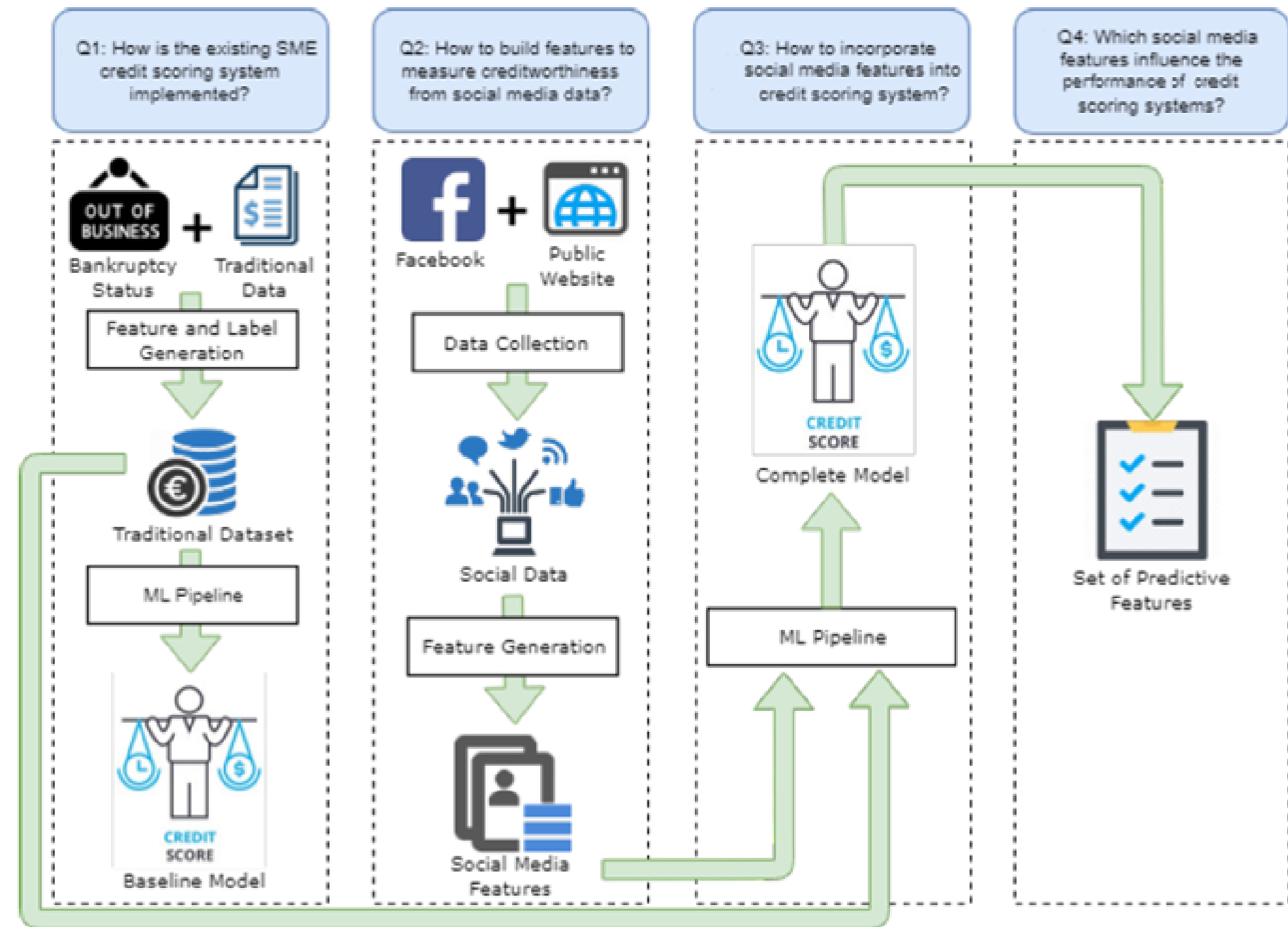
So that we also performed image sentiment analysis using convolutional neural networks. It perform analysis on user's visual posts and predict whether post content are positive negative or neutral.

Future Scope

1

Social credit score + traditional system

We can use this non-traditional system along with existing traditional credit scoring systems where we have user's bank and credit history to accurately predict better loan defaulters and trustworthiness of users.



Future Scope

2

Networkx

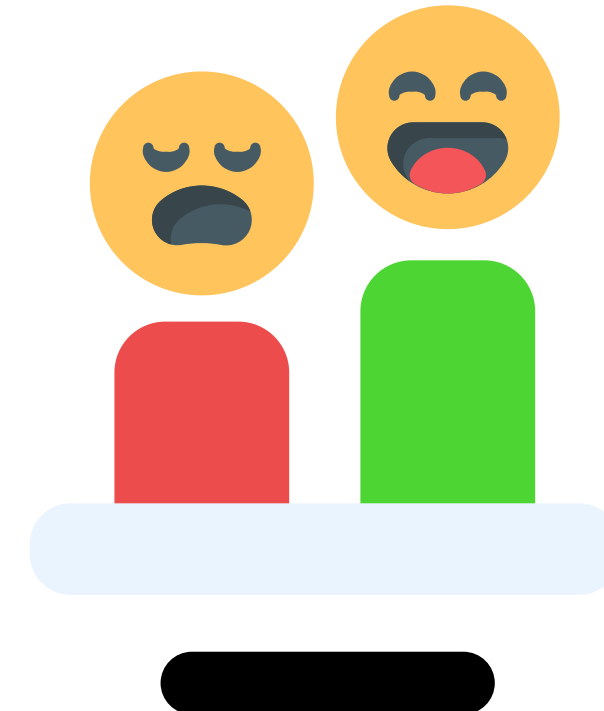
We can use Networkx libraries for better social network monitoring and analysis by the fact that who are potential friends and users in our user's network and how their score is also affecting the score of our user.



3

Emoji analysis

Make emoji analysis to create a more accurate score.





THANK YOU