Multimodal Popularity Prediction and Comparison of Top Beauty Brands – Lakme, L'Oreal & Maybelline Using Social Media.

By

Sharoz Siddique

Student ID: 2369480

Supervisor: Dr Mohammed Bahja

A thesis submitted to the University of Birmingham For the degree of MSc in Data Science [06 32255]

School of Computer Science

University of Birmingham, Birmingham, UK

September 2022

Table of Contents

Contents

List of Figures

List of Tables

List of Abbreviation

NLP	:	Natural language processing
RSME	:	Root mean square error
XGBoost	:	Extreme gradient boost
LBP	:	Local binary pattern
GCLM	:	Gray level cooccurrence matrix
GPU	:	Graphic processing unit
CNN	:	Convolutional neural network
API	:	Application program interface
PR	:	Precision and recall
ML	:	Machine learning
ANP	:	Adjective noun pairs
KNN	:	K-nearest neighbour
NB	:	Knowledge base
SMPD	:	Single program multi data
DNN	:	Deep neural network
VSCNN	:	Visual social convolutional neural network
DCNN	:	Deep convolutional neural network
ANN	:	Artificial neural network
SVM	:	Support vector machine
RTSP	:	Realtime streaming protocol
FCNN	:	Fully connected neural network
SVR	:	Support vector regressor
LBP	:	Local binary pattern
NLTK	:	Natural language tool kit
VADER	:	Valence aware dictionary for sentiment reasoning

Abstract

Brand-related user posts on social networks are expanding at an astounding rate, as industries compete for market share and try to boost their products' attractiveness by disseminating multimodal postings on social media platforms. But while some posts go viral, others are ignored. In our hypothesis, we describe a method for figuring out the factors that affect a post's popularity. According to our hypothesis, brand-related posts may be well-liked as a result of many indicators connected to factual data, sentiment, and visual imagery factors regarding the brand. We refer to this set of indications as engaging constraints. In our approach, we propose to use these indicators to anticipate the popularity of brand-related posts. Beauty related brands dataset is first scraped from Instagram and labelled as per the brand. Studies on a set of these Instagram posts shows that visual and textual features are complementary in predicting the popularity of a post. Concerning factual and visual imaging data taken into consideration, machine learning models have been used to determine the popularity using number of likes as the target feature. Additionally, as second part of this research, pre-trained NLP models have been applied to caption textual data to extract the sentiments from it. After determining the correlation between two findings, it was concluded whether it would be beneficial to combine them further in order to obtain superior results. The effectiveness of the suggested strategy in comparison to other studies has been shown by using RMSE as the criterion for evaluating the results of machine learning models. A comparison of the brands has also been established to determine which one became the most well-known based on the predicted values of likes taken as the target feature.

Keywords

Machine Learning – XGBoost; Linear Regression; Support Vector Regression; Random Forest.
NLP - Text Blob and Vader.

Acknowledgment

This thesis became a reality with the kind support and help of many individuals; I would like to extend my sincere thanks to all of them.

My deepest thanks to God for giving me the strength and ability to understand my studies and learn from them to complete this thesis.

I would like to express my sincere appreciation to my supervisor Dr. Mohammed Bahja and inspector Dr. Qamar Natsheh for giving me creative and comprehensive advice and for their guidance and support during the project. This research would not have been completed without their mentorship and advice during the semester.

Also, I would like to thank my parents and my wife for their continued encouragement, support and patience. They are the main source of inspiration and creativity in my life.

Finally, my thanks and appreciation goes to all those people who have willingly helped me out with their abilities, suggestions and support, you know who you are – Thank You!

CHAPTER I: Overview

1.1 Introduction

On social networking sites like Instagram, Twitter and Facebook, users or industries can create and share content (by like or commenting, for example). As a result, social media sites have become integral parts of our everyday lives and significant suppliers of social content. Due to the increasing expansion of social media content (i.e., texts, photos, audios, and videos), only a small part of online social content garners substantial attention and gains popularity, while the great majority either receives little attention or is completely ignored. Therefore, a lot of work has been put towards predicting the popularity of social media content, understanding its variance, and gauging its growth during the past few years [1]-[3]. This popularity reflects user preferences and offers insights into how users engage with digital content as well as how information spreads through social media networks.

Although, the problem of better understanding the factors that can have crucial impact on a post or image popularity is well explained in [4]-[7], but it still remains a challenging task. In this study, popularity prediction on Instagram is analysed to better understand the popularity features for an image. Popularity prediction can be significantly influenced by various factors (and features), such as visual content, aesthetic quality, user, post metadata, and time; therefore, considering all this multimodal information is crucial for an efficient prediction.

Additionally, it is difficult to choose a model that can reliably anticipate image popularity and make greater use of the different features that influence it. For an instance, machine learning models such as Support Vector or Decision Tree requires time and skill to adjust its hyperparameters with the use of highly structured data. With an extravagant popularity and performance XGBoost has been considered a check for evaluation on our dataset.

Accordingly, we study two crucial factors that will contribute to the popularity of an image, namely, visual content and social context, by analysing and structuring the dataset obtained from Instagram. Imagery features are in context of low level and high level; as per the studies held [8]-[11] LBP, GLCM and Gabor has been considered to be one of the best in terms of texture descriptor, characterisation and facial recognition respectively. Also, deep learning features though quite popular but it requires lots of data and high-performance GPUs. Convolutional neural networks, a class of deep learning techniques, perform feature extraction and categorization (CNN). CNN handles both feature extraction and classification based on a variety of images. When there is more data, deep learning algorithms perform better [12][13].

Likewise, by examining the three main categories of social features, we analyse the important role of context information connected with images; 1)Likes Count 2)Comments Count 3)Follows. Additionally, we have taken into account the textual information specific to the caption and extracted the sentiment in order to compare the results and determine whether there is a direct or inverse relationship between them. This is because the caption will always be a key element in drawing viewers and increasing the popularity of a post.

The researcher in the paper [34] has suggested that Textblob could be highly applicable in business intelligence and could do the sentiment analysis effectively. [35] Also showed some decent results with Text blob on sentiment classification.[36][37] has shown exceptional results with Vader model for sentiment analysis. In our approach, we have used both Textblob and Vader for sentiment analysis for each brand. Later, the results were compared with the final predicted results to draw a conclusion of whether it can have an impact on the predicted values.

1.2 Motivation

This research will use machine learning models to predict the post popularity using essential components which are visual, social and factual features for the top three competitive brands of the beauty industry. The project is driven by the following considerations:

1) Social media platforms have assimilated into our daily lives and are now important sources of social engagement. As social media information (texts, images, audio files, and videos) grows in volume, only a small portion of it receives significant attention and becomes popular, while the vast majority either receives little attention or is completely disregarded. As a result, much effort has been made to forecast the popularity of social media material.
2) In addition, since everyone's life now revolves around maintaining oneself and grooming, the beauty business is the one that receives the most attention and publicity. Additionally, the leading competitive brands are always concerned and want to compare themselves to their rivals. This research will benefit the market, but it will also have an impact on consumers' decisions to buy from the most popular and in-demand brands or industries.
3) The same model can be used across different industry type dealing with Instagram popularity. This research will illustrate the importance of introducing machine learning to the industrial environment. It can motivate the PR team to explore the use of ML in other case studies.
4) Most importantly, Multilabel dataset has not yet been introduced with multimodal popularity prediction approach. This research can open many research doors, from data scraping and labelling to making it structured and later merging of different types of features to improve the predictive performance of the model.

1.3 Research Question – Aims & Objective

1. Machine learning algorithms can assist industries in better understanding consumer preferences and trends for their products. How to leverage social media platforms to take advantage of the large market, in order to increase appeal, audience size, and perhaps revenue.

2. How a Multimodal approach which utilizes important aspects when it comes to feature that could improve the popularity prediction using the machine learning models.
   - What all High level and low level features of images shall be considered?
   - What social information shall be obtained from Json like no. of likes, comments etc.?
   - How textual information will be useful in sentimental analysis?

3. The study question aims to be resolved in this paper is the selection of appropriate models and setting their hyperparameters such that it aids in enhancing the performance because there are n number of machine learning models accessible for training our dataset.

4. Which pretrained model shall be applied for performing textual sentiment analysis & to build a conclusion on effect of sentiment on our resultant popularity prediction.

5. Merging all the above to conclude which brand gained the highest popularity based on the resultant predicted values?

1.4 Thesis overview:

The thesis begins with the literature review where have studied all the previous papers conducted in the field of Multimodal popularity prediction. As part of the study, we learned about several data scrape techniques and picked the one most suitable to carry out the operation of dataset extraction [43][44][45]. Instaloader has proven to be most effective for data scraping related to top 3 beauty brand (Lakme, L'Oreal and Maybelline). Feature extraction has been another crucial aspect and the deciding factor any model performance. Gabor Filter, Local Binary Pattern and Gray Level Co-occurrence matrix [8][9][10] has been proved the most efficient with exceptional accuracy of close 100% in the field of classification and face recognition and has been used for texture feature extraction of images, in our approach.

In the later stage of this report, will highlight the few well-known machine learning models and build a brief discussion supported with results to evaluate which model performs the best and provides the close precision between actual and predicted values. Number of likes and labels has been taken as the target feature which means it will need a multioutput regressor. Also, will discuss some of the important feature of XGBoost and how it is the best approach for our dataset. As will be building the regression model, RMSE has been taken as the performance metric to evaluate which model outperform the others.

Once the values has been predicted on the validation dataset for number of likes and labels for all the three models, will take the mean for the number of likes for each label (0:Loreal, 1:Lakme and 2: Maybelline). With the mean likes, will plot a graphical analysis to prove which brand gained the highest popularity.

Once the values has been predicted on the validation dataset for number of likes and labels for all the three models, will take the mean for the number of likes for each label (0: Loreal, 1: Lakme and 2: Maybelline). With the mean likes, will plot a graphical analysis to prove which brand gained the highest popularity. As part of extra research, we have taken caption textual data and implemented the sentiment analysis to get the polarity score. Furthermore, with the help of sentiment score will discuss, if the sentiment gets combined with our multimodal prediction model will benefit the result or not.

1.5 Challenges & Contribution:

The achievements attained with this research is as follows.

1) There is no dataset available for the beaty brands specifically the one's selected for study. Therefore, we have introduced a new dataset related to the popularity prediction for top 3 beauty brands (Lakme, L'Oreal and Maybelline) containing the images, json for social information and textual data, which could be beneficial to other research.

2) Multilabel dataset has not yet been introduced in the field of multimodal popularity prediction specifically for performing the regression task. This research can be useful for many researchers that could help them with steps from data scraping and labelling to making it structured and later merging of different types of features to improve the predictive performance of the regression model using multioutput approach to include label as one of the target features.

3) Beauty industry have been explored in the field of popularity prediction. This research could be a breakthrough and the same model can be used across different industry type dealing with Instagram popularity. This research will illustrate the importance of introducing machine learning to the industrial environment. It can help them to compare their predicted results with other brands and boast their appeal in the world of social media.

4) Previous studies [14][16][17] conducted for popularity prediction using machine learning model have achieved quite impressive results with RMSE. This approach was able to achieve exceptional performance with the newly introduced dataset, exploring the power of XGBoost in the area of machine learning, and achieving an extraordinary performance evaluated using RMSE metric of 0.03952, which to the best of my knowledge is the best achieved so far in this field.

1.6 Thesis Structure

Chapter II Literature review: This chapter discusses various previous studies conducted in this field of popularity prediction.

Chapter III Background: This chapter discuss about the necessary background about the Data Pre-processing for both popularity prediction and sentiment analysis and also discusses the machine learning models used.

Chapter IV Methodology (Dataset, Data Pre-processing, Machine Learning & NLP): In this chapter the proposed method including data pre-processing, machine learning models used for popularity prediction and NLP models utilized for sentimental analysis of textual data is explained in detailed with their respective outcomes.

Chapter V Result and Discussion: This chapter discusses about the experimental results and compares how XGBoost model is better than the other baseline models. Also, it shows which brand seems to have gained the higher popularity.

Chapter VI Conclusion, Limitation & Future work: This chapter discuss about the conclusion of our research, states the limitation and the future work that could be done in respect to it.

CHAPTER II: Literature Review

Introduction

The machine learning algorithms from earlier works that pertain to sentiment analysis and popularity prediction will be examined and discussed in this part. All of the studies looking at the performance of ML in the popularity prediction area are listed in Section 2.2. With an emphasis on pretrained classification models, the nature of the data, and overall performance, Section 2.3 discusses many sentiment classification strategies that have been studied in the literature.

2.1 Feature Learning in Multimodal Popularity Prediction

For instance, the study [2] showed that social indicators (such as the number of followers or the number of uploaded images) and image content (such as the gist, colour histogram, texture, colour patches, gradient, and deep learning features) have a substantial impact on image popularity.

The researcher employed in [4] that Adjective-Noun-Pairs (ANPs), along with context and user features (the concise tags, image descriptions that allow users to comprehensively detail their images in natural language), were used in the study to predict a succinct popularity score of social media images. The ontology, consisting in a collection of 3,244 ANP (Adjective-NounPairs), has been defined for visual features extraction. They showed that sentiment variables had a strong correlation with popularity and a lot of predictive value when combined with context features.

Almgren, K [3] used early popularity features, picture semantics, and social context to make predictions about an image's future popularity. They specifically took into account how popularity shifts over time by gathering data on the photograph within an hour after upload and monitoring its popularity (e.g., after a day, a week, a month). They also took into account the semantics of the image as well as the social environment (i.e., user information). Based on established natural language processing and clustering approaches, the semantic of the photos was recovered.

Allagwail, S. and their co researcher [9] has shown an exceptional accuracy of close to 100% used for classification and facial recognition with the consideration of image texture features extracted with The Local Binary Pattern, Gray Level Co-Occurrence Matrix, and the Gabor filter. Euclidean distances were further used for classification in the further process. Similarly, other paper [10][8] study uses three different texture picture datasets to offer a cluster-based feature selection methodology for adopting more discriminative subset texture characteristics. This involves extracting texture features using the Gabor filter, Local Binary Pattern, and Gray Level Cooccurrence Matrix (GLCM). Using KNN and NB classifiers, which were 99.9554% accurate and performant for the Kelberg dataset, research improved classification accuracy and performance.

Proposed study [11] shows that a texture feature extraction method will increase discrimination power for volumetric images. Tasks involving the classification of textured volumetric data may be handled by the approach. We combine two complimentary types of data—feature vectors derived from Local Binary Patterns (LBP) and approaches based on Gray-Level Cooccurrence Matrixes—to achieve this. They include details about the volumetric data's homogeneity, local anisotropy, and contrast as well as the visual pattern.

Researcher in [13] used SMPD2019 dataset considered features from four sources: visual content, text, users, and temporal-spatial information, and examine significant elements that may influence the popularity of postings in order to solve the problem of popularity prediction. Additionally, they suggested of combining features from different sources using the deep neural network (DNN) to forecast popularity. However, deep learning features though quite popular but it requires lots of data and high-performance GPUs.

A deep learning model called visual-social CNN (VSCNN) was proposed by researchers in [12] using a dataset of 432k images posted on Flickr. VSCNN predicts the popularity of a posted image by combining different types of visual and social features into a unified network model. It first learns to extract high-level representations from the input visual and social features by using two individual CNNs. To calculate the popularity score in the output layer, the outputs of these two networks are then combined into a single network.

Study in [14] have initially used a web crawler, specifically the Github Instagram-scraper, to gather a total of 1022 pictures from the Jeju Tourism Board's official Instagram account (@visitjeju.kr) (JTB). Later, utilised Microsoft Azure to extract the attributes of the image material. Each photograph has its own unique ID, posting time, date, and object category.Three different test set were created out of which Set-III stands out. Machine learning models were then used for popularity prediction.

2.2 Machine Learning in Multimodal Popularity Prediction

Several methods have been explored in order to classify the best machine learning model suitable for different types of datasets which when combined together results in better performance. Since the social media is exponentially growing platform it is necessary that the computer aided technology must be used for effective popularity prediction.

The researcher in [7] provides an approach to evaluate the overall sentiment of social media images related to a brand using both visual and textual indicators. Unlike previous studies, they focus on text that is embedded in an image rather than text that is presented alongside a picture. A machine learning classifier uses visual and textual information taken from two trained Deep Convolutional Neural Networks (DCNNs) to determine the sentiment of a picture. The method was evaluated on the "GfK Verein Dataset", and many machine learning techniques were compared. The performance of all classifiers is good, with F1-Scores ranging from 0.72 for NB to 0.79 for ANN.

Researcher in [14] proposed model training on three separate test sets, using four perspectives (visual, text, user, and temporal-spatial) to forecast the popularity of posts and train a DNNbased regression model to determine the final popularity score. The model significantly outperforms other models on the test Set-III (random partition). There were several models used, including Linear Regression, Random Forest Regression, GBM, however XGBoost performed the best with an RMSE of 0.1231

Proposed study [15] examined various image popularity prediction methods based on tag and visual features. Out of the 100 million public Flickr photos in the dataset used for Yahoo Flickr Creative, ten thousand images were chosen. They have demonstrated that, in unimodal learning, visual feature is less potent than tag feature by using various pre-trained deep learning models for visual feature extraction. With spearman's values of 0.260 and 0.263, respectively, Visual lib-linear and Visual lib-SVM exceed the other model tested.

Researcher in [16] a system that combines visual-textual characteristics with XGBoost for predicting popularity They combined the multi-modal characteristics and directly entered them into the XGBoost for popularity prediction after adopting a shape descriptor called Hu moment to extract the visual features from the image and on the other hand using one-hot encoder to encode the metadata of the posts. extensive research on the SMPD2019 dataset was done. XGBoost outscored competitors with an MAE of 1.195 and a spearman's rho of 0.7630 when many models were utilised, including Linear Regression, Random Forest, and Lasso.

Study [17] used one of China's biggest broadcast TV platforms, Jiangsu Cloud-media TV, provided the experimental data for this study. They obtained popularity time series and 11 static features by cleaning the RTSP (Real Time Streaming Protocol) packets from the video server and studying the EPG data. Created 10 folds from the dataset, of which 9 were used as the training set and 1 as the test set. Szabo-Huberman (S-H), Multivariate Linear, and MRBF models were compared to the prediction approach with the RMSE and R2 measurements in order to assess the prediction performance. 1.677 RMSE was attained with their approach.

2.3 NLP pretrained model Learning in Multimodal Popularity Prediction

Previously, several research have been conducted within this area analysing tweets for several analysis and conclusions using NLP models. According to [18], it considered tweets from leaders of political parties as a dynamical proxy to political programmes and ideas. Trained train a Fully-Connected Neural Network (FCNN) to recognise the political affiliation of a tweet and is s able to predict the origin of the tweet with a precision in the range of 71–75%, and the political leaning (left or right) with a precision of around 90%.

Researcher in [19] developed a natural language processing (NLP) based pre-processed data framework to filter tweets. Secondly, they incorporate Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) model concept to analyse sentiment, they also achieved 85.25% accuracy in sentiment analysis using NLP technique. According to [20] they automated AI-based observation framework to detect the emergence of public emotions and negativity in conversations. They evaluated the applicability of the framework using 29 928 social media conversations toward the much-debated topic of self-driving vehicles which will become increasingly relevant to smart cities. The patterns and transitions of citizens' collective emotions were modelled using the Natural Language Processing and Markov models while the negativity (toxicity) in conversations was evaluated using a deep learning-based classifier.

According to [21] they collected the tweets related to many trending topics, labelling them based on their content which is either malicious or safe. After a labelling process they extracted many features based on the language models using language as a tool. They also evaluate the performance and classify tweets as spam or not spam. Thus, our system can be applied for detecting spam on Twitter, focusing mainly on analysing of tweets instead of the user accounts. According to [22] they developed an approach that combines both sentiment analysis and classification. Thus, they were able to extract the topic in which users are interested. They have implemented algorithm using five lakhs of tweets and around a thousand of users.

2.4 Key Findings from Related Studies

A number of variables, including visual content, aesthetic quality, user and post metadata, can influence an image's popularity. Therefore, taking into account each of these aspects is crucial for effectively estimating image popularity. Additionally, the predictive model's effectiveness is quite important. Key techniques such as Feature selection, data pre-processing and machine learnings has been discussed.

2.4.1 Feature Selection

There may be several ways to extract information from images for the purpose of predicting popularity. Likewise, specifying the key features and limitations of each technique for its extraction from both visual and textual data is crucial.

2.4.1.1 Image Features

There are numerous methods for extracting image features, but the two most underlined and popular techniques are deep learning, where images are input directly into the model, which then outputs a set of features; and other technique is by using scikit-image, which is a collection of image processing algorithms.

- Using Scikit-Image

For features, there are two basic sorts of procedures: transform-based methods and statistical methods. Grey Level Co-occurrence matrix(GLCM) and Local Binary Pattern (LBP) are examples of structural features based on topological and geometric aspects. The spatial frequency characteristics of the fluctuations in pixel intensity, like Gabor, are used by transform-based texture analysis techniques to transform the image into a new form [23]. According to the studies conducted [8,9,10,11], LBP, GCLM, and Gabor were shown to be among the best in characterising textures, identifying faces, and descriptors of texture, respectively. Our proposed study has therefore considered all the three elements LBP, GLCM and Gabor for our experiment. Scikit image is used as an image processing toolbox containing modules for such operations.

- Using Deep Learning

Several Deep learning models such as visual-social CNN(3 layer model) [12], Deep Neural Network(DNN) [13] and Deep Convolutional Neural Networks (DCNNs) [7] are used where image are fed which outputs a set of features. However, the deep learning models requires a huge amount of data and high-performance GPUs which could be the limitation for our study as the dataset used may not be suitable for this technique.

2.4.1.2 Post Metadata Features

Every post includes metadata, which when combined with an image's set of features can help a model's performance. This metadata alone has the ability to predict how popular a post will be. Contextual information, such as the number of comments and descriptions, dimensions of image, social information, such as the number of likes and who they are followed by, as displayed in the study [24]. Additionally, factors like popularity variations over time (days, weeks, and months) [3] have been looked into. [13] demonstrated that the number of likes and comments complement one another and are an excellent indicator of popularity. According to our proposed study, the essential features for metadata analysis include a combination of contextual features, such as the number of comments, image dimensions (Width, Height), and social information, such as the number of likes and follows.

2.4.2 Model Selection

Without being specifically programmed to do so, machine learning (ML), a subset of artificial intelligence, enables software systems to improve their propensity for outcome prediction. ML algorithms forecast new output values using historical data as input. ML algorithms can now analyse images in the same way as our brains do when dealing with visual data. ML algorithms typically contain predefined pipelines or phases for learning from data. First of all, in order to learn and predict outcomes with great accuracy, ML requires a sizeable volume of high-quality data. To ensure that ML image processing works properly, images must be properly processed, tagged, and generalised. There are several libraries and frameworks for image processing that can help with this. The field of machine learning has undergone a revolution because to models like XGBoost.

- XGBoost Regressor:

The strength of this potent method resides in its scalability, which promotes rapid memory utilisation and parallel and distributed computing for fast learning. XGBoost is a technique for group learning. It might not always be enough to rely solely on a single machine learning model's output. It has a methodical approach of combining the prediction capacity of various learners is provided by ensemble learning. The resultant is a single model which gives the aggregated output from several models. Boosting, Regularisation and handling sparse data are one of the key features of XGBoost. With Multimodal Popularity prediction XGBoost has always topped the position by its performance [14][25][16][26][27], displaying the best results with respect to metric of RMSE, MAE respectively.

- Multioutput Regressor

Most of the machine learning models strategy consist of fitting one regressor per target. Multioutput Regressor is a simple strategy for extending regressors that do not natively support multi-target regression. With our approach considering the label and number of likes as two interrelated outputs, as it is important to know the predicted likes is associated with which label or brand. Hence with each machine learning model that doesn’t support multi target, multioutput regressor has been accommodated with it.

- Random Forest Regressor:

A popular algorithm for classification and regression issues is the supervised machine learning technique known as random forest. It can handle binary, continuous, and categorical data. It creates decision trees from various samples, relying on their majority for categorization and average for regression. One of the key characteristics of the Random Forest Algorithm is its ability to handle data sets with continuous variables, such as those used in regression analysis. It maintains diversity as all the attributes are not considered while making each decision tree though it is not true in all cases. Also, it is immune to the curse of dimensionality. Since each tree does not consider all the attributes, feature space is reduced. [14][16] and proposed method shows some decent numbers for our performance metric.

- Support Vector Regressor (SVR):

With SVR the aim is to find a function that approximates mapping from an input domain to real numbers on the basis of a training sample. when moving on with SVR, is to basically consider the points that are within the decision boundary line. The best fit line is the hyperplane that has a maximum number of points. SVR acknowledges the presence of non-linearity in the data and provides a proficient prediction model. The proposed method and [14][16][25] displays a decent and a competitive performance with respect to random forest and linear regression.

- Linear Regression (LR):

Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent and independent variable. Linear Regression is of two types: Simple and Multiple. Simple Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable[14][25]. Whereas, In Multiple Linear Regression there are more than one independent variables for the model to find the relationship.

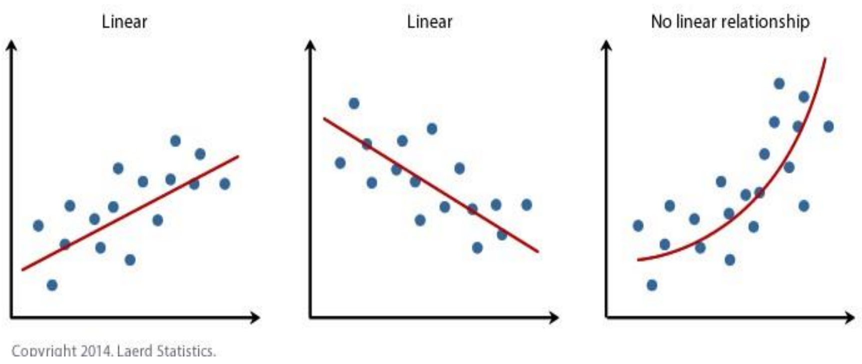


Fig 1.2: Linear Regression

2.4.3 Knowledge Gap

- Most of the popularity prediction studies only consider single modality (say, only text or image) into account, which limits their performance of popularity prediction.[28] has first pre-trained a deep model that predicts the popularity of recipes based on each single modality. The popularity of a recipe is based on its content features which they have predicted even before users post their comments on the recipe.

- To the best of my knowledge, no work has been performed in building a model that takes a multilabel dataset and then performs popularity prediction with multioutput regression, instead a classification model was built [29]. In this framework, this framework, a video dataset containing natural disasters is used for multi-label classification. But again, though multimodal it doesn't have to do anything with popularity prediction.

- Most of them focused on visual content features extracted from deep learning techniques, rather than considering other texture features like LBP, GLCM, Gabor which has a high success rate as a texture descriptor and have proved to have an accuracy of close 100% in the field of classification and face or object recognition [8][9][11].

- To the best of my knowledge, model implemented so far were able to achieve an RMSE mostly in the range of 0.1231 [14] and MAE in the range of 1.195 [16]. Our approach aims for better and efficient performance of model by best feature selection and structuring of data.

- Our study has introduced a new dataset with around 4K images and JSON for context features and Text file for textual data scraped from Instagram using instaloader, for the top 3 beauty brands – L'Oreal, Lakme and Maybelline.

CHAPTER III: Background

Introduction

This section aims to provide the high-level view of the algorithms used in this project to help with the understanding of upcoming sections. We will elaborate on the techniques and libraries used as part of the Data Pre-processing in section 3.1. In section 3.2 we aim to discuss about the functions used to extract the texture features from image and contextual features from json. Further in section 3.3, will discuss about the techniques and pretrained model utilized to get the sentiment from textual caption data, so that it can be later used with prediction model to build a conclusion. Section 3.4 discusses about the vital machine learning models used in our research with performance metric used for evaluation in the following section 3.5.

3.1 Data Pre-processing

To ensure that the data is structured, cleaned, and processed in such a way that when fed to the machine learning model elevates the performance. Below are some of the popular techniques used for pre-processing and cleaning the data:

The first step to data pre-processing was label the dataset (First part of dataset: Images and Json), the process of which will be discussed in the later section. In general, ML datasets contain a variety of features with varied strengths, ranges, and units. This is a serious challenge because some machine learning algorithms are quite sensitive to these features. This is where the concept of feature scaling comes in play. It’s an essential part of the data pre-processing stage.

- Scikit Learn:

Many Python modules offer reliable implementations of many machine learning methods. One of the most well-known is Scikit-Learn, a package that offers effective versions of many popular algorithms. Scikit-Learn is distinguished by a clear, consistent, and streamlined API in addition to extensive and helpful online documentation. This uniformity has the advantage that moving to a different model or method is simple after you grasp Scikit-fundamental Learn's usage and vocabulary for one type of model.

- Scaling:

Consider a dataset with multiple features, there might be a case where the scales or values of these features are very different, one might have higher scales and the other not. Therefore, there is a chance that higher weightage is given to features with higher magnitude. This will impact the performance of the machine learning algorithm and we do not want our algorithm to be biased towards one feature.

Therefore, we scale our data before employing a distance-based algorithm so that all the features contribute. Distance algorithms like KNN, K-means, and SVM are most affected by the range of features. Hence, Scaling is mostly applicable to such algorithms.

	Student	CGPA	Salary '000
0	1	3.0	60
1	2	3.0	40
2	3	4.0	40
3	4	4.5	50
4	5	4.2	52

Fig 1.1: Before Scaling

	Student	CGPA	Salary '000
0	1	-1.184341	1.520013
1	2	-1.184341	-1.100699
2	3	0.416120	-1.100699
3	4	1.216350	0.209657
4	5	0.736212	0.471728

Fig 2.2: After Scaling

- Normalisation:

A scaling technique called normalisation shifts and rescales values so that they fall between the ranges of 0 and 1. Additionally called Min-Max scaling.

$$X' = \frac{X - XMIN}{XMAX - XMIN}$$

The feature's maximum and minimum values are indicated here by the letters Xmax and Xmin, respectively. The numerator will be zero when the value of X is the lowest number in the column, hence X' will be 0. The numerator is equal to the denominator when the value of X is the highest value in the column, however, and X's value is therefore 1. The value of X' is between 0 and 1 if the value of X is between the lowest and maximum value. The feature's maximum and minimum values are indicated here by the letters Xmax and Xmin, respectively. The numerator will be zero when the value of X is the lowest number in the column, hence X' will be 0. The numerator is equal to the denominator when the value of X is the highest value in the column, however, and X's value is therefore 1. The value of X' is between 0 and 1 if the value of X is between the lowest and maximum value.

- Data Preprocessing for Sentiment Analysis

The second part of the dataset is the textual data. Before using any sentiment analysis approaches, data must first go through a pre-processing stage to make sure that only relevant words are included in the classification process.

i. Tokenisation: A frequent task in natural language processing is tokenization (NLP). Both conventional NLP techniques like the Count Vectorizer and advanced deep learning-based frameworks like Transformers use it as a crucial step. Tokenization is the process of breaking up a long block of text into tokens. Tokens in this context can be words, characters, or subwords. The three main categories of tokenization are word, character, and subword (n-gram characters) tokenization [30].

ii. Porter stemmer: Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. For instance, stemming will reduce words like "Imagine", "Imagination" and

"Imaginative" to "Imagin"[31]. Porter presented a simple algorithm for stemming English language words developed in 1968.

iii. Remove Stop words: A language's stop words are a group of frequently used terms. Stop words in English include "a," "the," "is," "are," and others. Stop words are frequently employed in text mining and natural language processing (NLP) to get rid of words that are so frequently used that they don't actually contain any meaningful information. For instance, if your search query is "what is a imaginary world" in the context of a search system, you want the search system to prioritise surfacing documents that discuss imaginary world over documents that discuss what is a [19]. You can achieve this by keeping a list of stop words (which can be manually or automatically curated) and preventing the analysis of any words on your stop word list.

iv. Remove punctions, emojis and url: Emojis, punctuation, and urls are removed from the text. Such terms are frequently used in documents but have no sentimental meanings and can affect subsequent calculations of terms' weights.

3.2 Feature Representation

There are numerous methods for extracting image features, but the most underlined and popular techniques is by using scikit-image, which is a collection of image processing algorithms.

3.2.1 Local Binary Pattern (LBP)

Local Binary Patterns, or LBPs for short, are a texture descriptor that was first introduced in 1993 and gained popularity because to the work of Ojala et al. in their 2002 study. A local representation of texture is computed via LBPs. This local representation is created by comparing each pixel with the pixels in its immediate neighbourhood. The image must be converted to grayscale before creating the LBP texture descriptor. We choose an area of size r around the central pixel for each pixel in the grayscale image. In the output 2D array, which has the same width and height as the input image, an LBP value is then computed for this centre pixel [8][9][11].

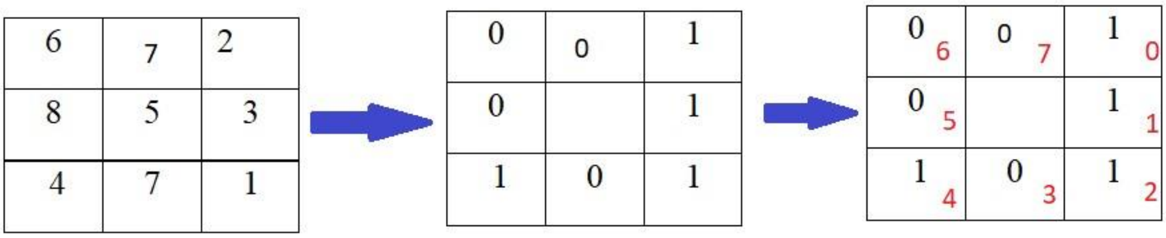


Fig 3.1: Constructing a LBP is to take the 8 pixel neighbourhood surrounding a center pixel and threshold it to construct a set of 8 binary digits.

0	0	0	1	0	1	1	1

1+2+4+32 = 39

Fig 3.2: Taking 8-bit binary neighbourhood of the center pixel and converting it into a decimal representation.

Fig 3.3: An example of computing the LBP representation (right) *from the original input image* (left).

3.2.2 Gray-level Co-occurrence Matrix (GLCM)

A statistical method of examining texture that considers the spatial relationship of pixels is the gray-level co-occurrence matrix (GLCM), also known as the gray-level spatial dependence matrix. The GLCM functions characterize the texture of an image by calculating how often pairs of pixels with specific values and in a specified spatial relationship occur in an image, creating a GLCM, and then extracting statistical measures from this matrix. (The texture filter functions, described in Calculate Statistical Measures of Texture cannot provide information about shape, that is, the spatial relationships of pixels in an image)[10][11].

Statistic	Description	Formula
Contrast	Measures the local variations in the gray-level co-occurrence matrix.	$Contrast = \sum_{i,j=0}^{N-1} P_{ij}(i-j)^2$
Correlation	Measures the joint probability occurrence of the specified pixel pairs.	$Correlation = \sum_{i,j=0}^{N-1} P_{ij}\frac{(i-\mu)(j-\mu)}{\sigma^2}$
Energy	Provides the sum of squared elements in the GLCM. Also known as uniformity or the angular second moment.	$Energy = \sum_{i,j=0}^{N-1} \left(P_{ij}\right)^2$
Homogeneity	Measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal.	$Homogeneity = \sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+(i-j)^2}$

Table 1: GLCM Properties

where:

Pij = Element i,j of the normalized symmetrical GLCM

N = Number of gray levels in the image as specified by Number of levels in under Quantization on the GLCM texture page of the Variable Properties dialog box.

μ = The GLCM mean (being an estimate of the intensity of all pixels in the relationships that contributed to the GLCM), calculated as:

σ ^{2} = the variance of the intensities of all reference pixels in the relationships that contributed to the GLCM, calculated as:

$$\sigma^2 = \sum_{i,j=0}^{N-1} P_{ij} (i - \mu)^2$$

Fig 3.4: GLCM Contrast

3.2.3 Gabor Filter

Gabor filtering for image textural analysis has been introduced by Daugman [32]. The success of Gabor filters in this field is due to their aptitude to model the response of simple cortical cells in the visual system. A 2D Gabor filter can be thought of as a complex plane wave modulated by a 2D Gaussian envelope and can be expressed in the spatial domain as:

$$G_{\theta,f,\sigma_1,\sigma_2}(x, y) = \exp\left[\frac{-1}{2}\left(\frac{x'^2}{\sigma_1^2} + \frac{y'^2}{\sigma_2^2}\right)\right] \cos(2\pi f x' + \varphi)$$
$$x' = x \sin\theta + y \cos\theta$$
$$y' = x \cos\theta - y \sin\theta$$

where f is the spatial frequency of the wave at an angle θ with the x axis, σ1 and σ2 are the standard deviations of the 2D Gaussian envelope, and ϕ is the phase.

Frequently in textural analysis applications, and also in this case, the Gaussian envelop is symmetric, so we have σ = σ1 = σ2. A Gabor filter is suited to obtain local frequency information in a specific orientation (given by θ), which is directly related with image contours [33].

3.3 Sentiment Analysis

Sentiment analysis, or opinion mining is a technique for discovering people's opinions and ideas within a document on a specific subject. It can also be used to rate a product or event, as with movie reviews, and is typically used to categorise customer comments as positive, neutral, or negative.

The following subsections explain some of the popular pretrained text classification techniques. Section discusses popular pretrained models the Vader (3.3.1) and TextBlob (3.3.2) techniques.

- NLTK Library

It is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

- Googletrans

It is a free python library that uses Google Translate API. In this article, we explain how to employ the library to translate strings as well as data frames in Python. Fast and reliable - it uses the same servers that translate.google.com uses, Auto language detection, Bulk translations and Customizable service URL are few of the important feature of google translation. We have used translator feature of the module

3.3.1 TextBlob

A Python library for Natural Language Processing is called TextBlob (NLP). Natural Language ToolKit (NLTK) was a tool that TextBlob actively employed to complete its tasks. The NLTK library enables users to do categorization, classification, and a variety of other tasks while providing quick access to a large number of lexical resources. TextBlob is a straightforward library that provides intricate textual analysis and processing.

A sentiment is identified by its semantic orientation and the force of each word in the sentence for lexicon-based approaches. This calls for a pre-defined dictionary that divides words into negative and positive categories. A text message will typically be represented by a bag of words. Following the individual scoring of each word, the ultimate sentiment is determined by performing a pooling operation, such as averaging all the sentiments.

TextBlob returns a sentence's polarity and subjectivity. The polarity scale is [-1,1], where -1 represents a negative emotion and 1 represents a good emotion. Negative words turn the polarity around. Semantic labels in TextBlob facilitate detailed analysis. Emoticons, exclamation points, emoticons, etc. are a few examples. The range of subjectivity is [0, 1]. Subjectivity measures how much factual information and subjective opinion are present in the text. The content contains personal opinion rather than factual information due to the text's heightened subjectivity. One other setting for TextBlob is intensity. TextBlob uses the "intensity" to determine subjectivity. Whether a word modifies the next word depends on its intensity. Adverbs are used as modifiers in English, such as "extremely good."

The researcher in the paper [34] has suggested that Textblob could be highly applicable in business intelligence and could do the sentiment analysis effectively. [35] Also showed some decent results with Text blob on sentiment classification.

3.3.2 Vader

A model for text sentiment analysis called VADER (Valence Aware Dictionary for Sentiment Reasoning) is sensitive to both the polarity (positive/negative) and intensity (strong) of emotion. It may be used right away on unlabelled text data and is included in the NLTK package. The VADER sentimental analysis uses a dictionary that converts lexical data into sentiment scores, which measure the intensity of an emotion. By adding the intensity of each word in a text, one can determine the sentiment score of that text.

For instance, the words "love," "enjoy," "glad," and "like" all express a good feeling. Additionally, VADER is wise enough to comprehend the underlying meaning of these terms, such as the negative connotation of the phrase "did not love." Additionally, it is aware of the significance of capitalization and punctuation.

- Polarity classification

Instead of making an effort to distinguish between a sentence's subjectivity, objectivity, or truthfulness. it is simply interested in whether the text communicates a positive, negative, or neutral opinion.

- Coarse analysis

It won't do a fine-grained analysis to figure out how positive or negative anything is. In other words, it doesn't try to guess the number of stars a reviewer gave, merely if the review was positive or negative.

[36] shows that the VADER Sentiment Analyzer was an effective choice for sentiment analysis and classification of Twitter data. VADER easily and quickly classified huge amounts of data. Also, [37] study shows that VADER outperforms Naive Bayes, Maximum Entropy, and Support Vector Machine (SVM) algorithms with Classification Accuracy of 0.96.

3.4 Machine Learning

Machine learning is a group of statistical tools and techniques that can use experiences and previous occurrences to train models to predict future events. Machine learning can improve decision-making and boost the accuracy of predictions by studying the data[38].

Machine learning techniques are divided into three learning types:

Supervised Learning: Using labelled examples and new data, supervised machine learning algorithms use what they have learnt in the past to forecast future events. An inferred function to forecast output values is produced by the learning algorithm by examining a known training dataset. After sufficient training, the system is capable of providing targets for any new input. To identify errors and correct them, the model can be modified by comparing its output with the proper, intended output[39].

Unsupervised Learning: When training data is neither categorised nor labelled, unsupervised machine learning techniques are utilised. Unsupervised learning investigates how systems might infer a function from unlabelled data to describe a hidden pattern. The system can never be guaranteed that the output is correct. Instead, it infers what the result should be based on datasets [39].

Semi Supervised: The supervised and unsupervised learning models are combined in this: Some of the training data is labelled, but some will lack labels [39].

As was indicated before in the previous section, classification and regression issues can both be resolved via supervised learning. The research's nature dictates that this part will concentrate on regression models. The regression category includes several machine learning models.

Some of the well-known classification regression models are explained in the following subsections.

3.4.1 XGBoost Model:

XGBoost is a perfect blend of software and hardware capabilities designed to enhance existing boosting techniques with accuracy in the shortest amount of time. It might not always be enough to rely solely on a single machine learning model's output. It has a methodical approach of combining the prediction capacity of various learners is provided by ensemble learning. The resultant is a single model which gives the aggregated output from several models. Boosting, Tree Pruning and handling sparse data are one of the key features of XGBoost.

- Tree Pruning:

Tree pruning is a machine learning strategy that involves removing nodes that don't improve the classification of nodes in order to reduce the size of regression trees. The purpose of pruning a regression tree is to avoid the training data becoming overfit. Cost Complexity or Weakest Link Pruning, which internally uses mean square error, k-fold cross-validation, and learning rate, is the most effective pruning technique. Up to the chosen max depth, XGBoost builds nodes (also known as splits) and begins backward pruning until the loss is below a predetermined threshold [40].

- Sparsity Aware Split Finding:

It frequently happens that the data we collect is sparse (has many empty or missing values) or that it becomes sparse as a result of data engineering (feature encoding). A default direction is given to each tree in order to make it aware of the patterns of data sparsity. In order to reduce training loss, XGBoost handles missing data by putting them in the default direction and obtaining the best imputation value. The technique is optimised in this case to visit only missing values, which makes it run 50x faster than the naive approach [40].

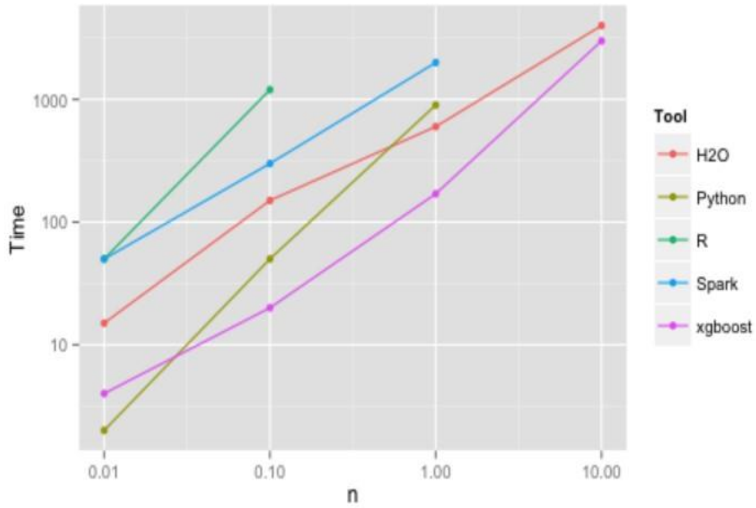
- Built-in Cross-validation:

A statistical technique called cross validation is used to evaluate machine learning models on unseen data. When the dataset is small, it is helpful because it avoids overfitting by not using an independent sample (holdout) from the training data for validation. By limiting the amount of training data, we risk damaging the hidden features and patterns that could cause more errors in our model. This is comparable to the scikit-learn library's cross val score functionality [40].

- DMatrix

It is an internal data structure used by XGBoost which is optimized for memory efficiency and training speed. We need to transform our numpy array of data using DMatrix so that it can be later utilized in dtrain and dtest parameters of its inbuilt functions [40].

Model tuning in XGBoost can be implemented by cross-validation strategies like GridSearchCV.



Benchmark Performance of XGBoost (source)

*Fig 4: O*bjective benchmark comparison of XGBoost with other gradient boosting algorithms trained on a random forest with 500 trees, performed by Szilard Pafka.

3.4.1.1 XGBoost Parameters:

- Learning rate alias: eta [default=0.3] [range = 0 to1]

Updates use step size shrinkage to prevent overfitting. We can directly get the weights of new features following each boosting step, and eta lowers the feature weights to make the boosting process more conservative [41].

- max_depth [default=6] range [0, ∞]

Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit. 0 indicates no limit on depth. Beware that XGBoost aggressively consumes memory when training a deep tree. exact tree method requires non-zero value [41].

- min_child_weight [default=1]

Minimum sum of instance weight (hessian) needed in a child. If the tree partition step results in a leaf node with the sum of instance weight less than min_child_weight, then the building process will give up further partitioning. In linear regression task, this simply corresponds to minimum number of instances needed to be in each node [41].

- eval_metric [default according to objective]

Evaluation metrics for validation data, a default metric will be assigned according to objective (rmse for regression, and logloss for classification, mean average precision for ranking) [41].

3.4.2 Support Vector Regressor

A non-probabilistic model called the supervised machine learning (SVM) can be utilised to address classification and regression issues. The decision boundary used by SVM to distinguish between classes is known as a hyperplane. Based on the data point's position in reference to hyperplane, each data point is either classed as a 1, or a 1 (above or below) [48].

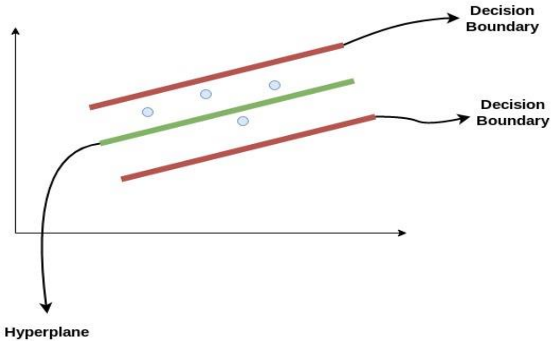


Figure 4.1: Support Vector Regressor (Ref: https://www.analyticsvidhya.com/blog/2020/03/support-vector-regressiontutorial-for-machine-learning/)

These are the terms used in SVR Regressor

The function called a kernel is used to transform lower-dimensional data into higherdimensional data.

Hyper Plane: In SVM, this is essentially the boundary between the several data classes. Even though we will define it in SVR as the line that will assist in predicting the continuous value.

Boundary line: Other than Hyper Plane, two lines in SVM generate a boundary. The support vectors may be outside or on the boundary lines. The two classes are divided by this line.

Assuming our hyperplane is a straight line going through the Y-axis. Equation is as follows.

$$Wx + b = 0$$

So the equation of the boundary line can be stated as

$$Wx + b = +e$$

$$Wx + b = -e$$

Thus, stating the fact that for for any linear hyper plane the equation that satisfy our SVR is:

$$e \leq y - Wx - b \leq +e$$

3.4.3 Random Forest Regressor:

Prior to discussing about random forest, it's critical to comprehend decision tree method. A supervised learning model called a decision tree can resolve classification and regression problems. By deriving branching rules from features, decision trees can forecast class labels [57], [64]. Splitting decisions are discovered by computing Gini impurity as follows:

$$IG(P) = 1 - \sum Pi\ C\ 2\ \mathrm{C}$$

- number of classes

$i \in \{1,2, . . . , C)$.

Pi denote the fraction of points at P that belong to class C.

Decision Tree may result into overfitting of the data. To overcome this Random Forest was introduced

Random Forest Regression combines several random decision trees, each of which has been trained on a subset of data. The algorithm is more stable and less erratic when many trees are used. The random forest regression approach, which performs well for big and most types of data, is a frequently used model [48][49[50].

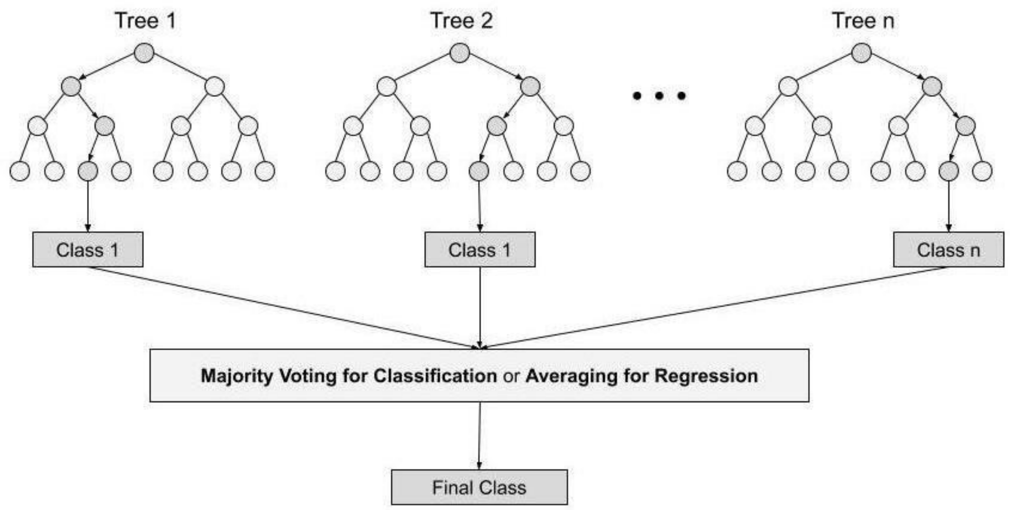


Fig 4.2: Random Forest

Features:

- Boosting:

Generally, refers to improving performance. A weak hypothesis or weak learners are turned into strong learners using the sequential ensemble learning technique known as "boosting" in machine learning. This increases the model's accuracy. A classifier that has a weak connection with the real value is technically referred to as a weak learner. Because of this, in order to create a model that combines the predictions of weak learners to create a strong learner, we use the boosting technique.

- Ensemble Learning:

Ensemble learning is a process in which decisions from multiple machine learning models are combined to reduce errors and improve prediction when compared to a Single ML model. Then the maximum voting technique is used on aggregated decisions.

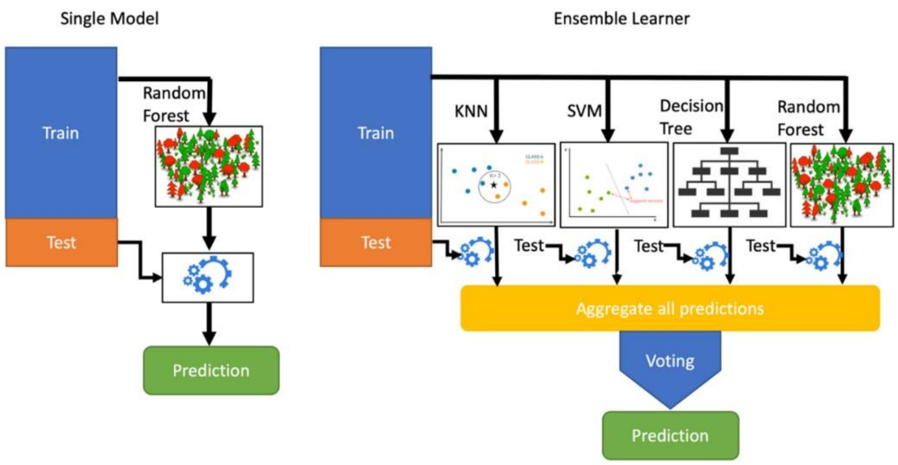


Fig 4.3: Single Model Prediction vs Ensemble Learning (Ref: https://medium.com/sfu-cspmp/xgboost-a-deep-dive-intoboosting-f06c9c41349)

3.4.4 GridSearchCV:

We pass on a parameter's dictionary to the function and compare the cross-validation score for each combination of parameters (many to many) in the dictionary and return the set having the best parameters [40].

Parameters:

- Estimator: estimator object

This is assumed to implement the scikit-learn estimator interface. Either estimator needs to provide a score function, or scoring must be passed [40].

- Param_grid: dict or list of dictionaries

Dictionary with parameters names (str) as keys and lists of parameter settings to try as values, or a list of such dictionaries, in which case the grids spanned by each dictionary in the list are explored. This enables searching over any sequence of parameter settings [40].

- Scoring: str, callable, list, tuple or dict, default=None

Strategy to evaluate the performance of the cross-validated model on the test set [40].

- Verbose: int

Controls the verbosity: the higher, the more messages [40].

i. >1: the computation time for each fold and parameter candidate is displayed;
ii. >2: the score is also displayed.
iii. >3: the fold and candidate parameter indexes are also displayed together with the starting time of the computation.

3.4.5 Performance Metric:

- Root Mean Squared Error (RMSE):

When training regression or time series models, one of the most widely used metrics to measure how accurately our forecasting model's predicted values compare to the actual or observed values is the root-mean-square error, or RMSE.

It measures the error in our predicted values when the target or response variable is a continuous number. For example, when using regression models to predict a quantity like income, sales value/volumes, demand volumes, scores etc. In our case, number of likes is the target feature which will be used for prediction. Thus, RMSE is a standard deviation of prediction errors or residuals. It indicates how spread out the data is around the line of best fit To calculate the RMSE the formula is.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2}$$

Where:

- ∑ is the summation of all values
- f is the predicted value
- is observed or actual value
- (fi — oi) 2 are the differences between predicted and observed values and squared
- N is the total sample size

To summarize our learnings on RMSE:

- RMSE is the standard deviation of the residuals.
- RMSE indicates average model prediction error.

CHAPTER IV: Multimodal Popularity Prediction - Methodology

Introduction

This chapter discusses the methodologies and tools used to train dataset containing visual and contextual data to predict the popularity of all three brands (Lakme, L'Oreal and Maybelline) based on likes and label (target features) and later build a comparison between them. In addition, build a sentiment classifier on the textual data for all the three brands and discuss at the end whether it can have an impact on improvement of popularity prediction or not. Section 4.1 describes the source of the dataset used in this research and methods used to scrape the dataset. Section 4.2 lists all the tools used in this project. Section 4.3 discusses the data pre-processing schemes used in this research and analyses the preprocessed data for insights. Sections 4.4 and 4.5 train machine learning and implementation of Machine Learning and pretrained NLP models.

4.1 Dataset Description

As businesses battle for market share and attempt to increase the attractiveness of their products by propagating multimodal messages on social media platforms, brand-related user posts on social networks are growing at an incredible rate. While some posts gain a lot of attention, others are overlooked. We propose a method for identifying the elements influencing a post's popularity. Our thesis states that a number of indicators associated to factual information, visual imagery (First part of the dataset definition) and Sentiment (second part of the dataset definition) aspects about the brand may result in brand-related posts being well-liked. This group of clues is what we refer to as engaging constraints. In our strategy, we suggest using these indicators to predict how popular brand-related postings will be.

In our thesis, we have leveraged Instagram, the most popular social media network, to apply our strategy. Scraped the dataset to include textual (.txt), contextual (.json), and visual (images -.jpg) files. The Json and Images combinely form the first part of the datset used for popularity prediction, and Json were used as the second part of the dataset used for sentiment analysis. Later, both of the result were compared and discussed to build a conclusion, whether both are co-related or not. With our approach, a new dataset was introduced containing 4295 unlabelled datapoints inclusive of all the three brands. The data can be scraped using the Application Programming Interface (API) of Instagram. Another way to do it is by python programming language using the Instalooter [42], Beautiful Soup Library [43] and Instaloader (Or an application) [44].

4.1.1 Web Scraping for Dataset.

In our study, we proposed to use instaloader for data scraping from Instagram. Alexander Graf created the Python library "Instaloader" to extract desirable post URLs. Based on the hashtag for a brand, URLs are chosen. The technique permits downloading posts for the dataset but restricts the total number of posts. Instaloader does self-pre-processing to clean up the dataset and eliminate redundancy. About 1200 photos were retrieved from the dataset for each brand using the Instaloader script. Using brand-specific hashtags and public profiles, such as

#Maybelline and Maybelline (official public account), it is possible to extract images posted by brands for advertising and to gain attention for their products with a high degree of probability of creating a useful dataset. Several attributes were gathered via Instaloader, including followers/following counts, comments and likes count, dimension of an image etc.

Process carried out for data scraping from Instagram.

1) Open the command prompt using the cmd command in windows Run.
2) Change the directory to desired folder using the cd command
   (Example: cd C:\Users\Sharoz\Dowloads\Dataset)
3) start https://github.com/instaloader/instaloader.[44]45]
4) curl-OL https://github.com/instaloader/instaloader/releases/download/v4.9.1/instaloaderv4.9.1-windows-standalone.zip (curl stands for client URL, a command line tool that developers use to transfer data to and from a server)
5) tar -xvf instaloader-v4.9.1-windows-standalone.zip (tar command helps create, extract, and list archive contents)
6) ren instaloader.exe ins.exe (Renaming the .exe file from instaloader to ins)
7) move instaloader.exe.md5 C:\path
8) move ins.exe C:\path
9) start C:\path (Create a path directory in your C drive)

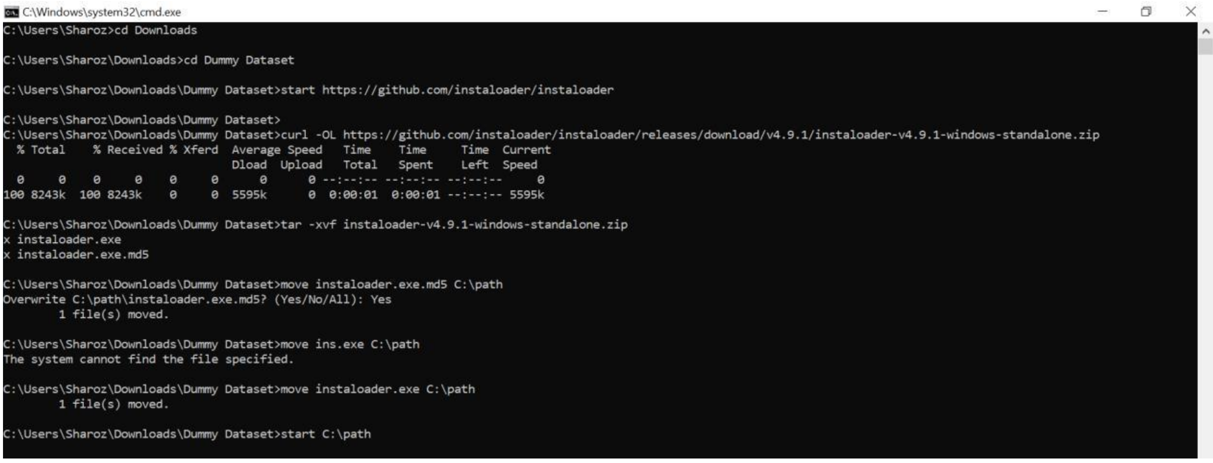


Fig 5.1: commands executed on command prompt for data scraping.

10) Open the Application and run the command (Example: ins profile Maybelline) ins [--comments] [--geotags]

    [--stories] [--highlights] [--tagged] [--igtv]

    [--login YOUR-USERNAME] [--fast-update] profile |

    "#hashtag" | %%location_id | :stories | :feed | :saved

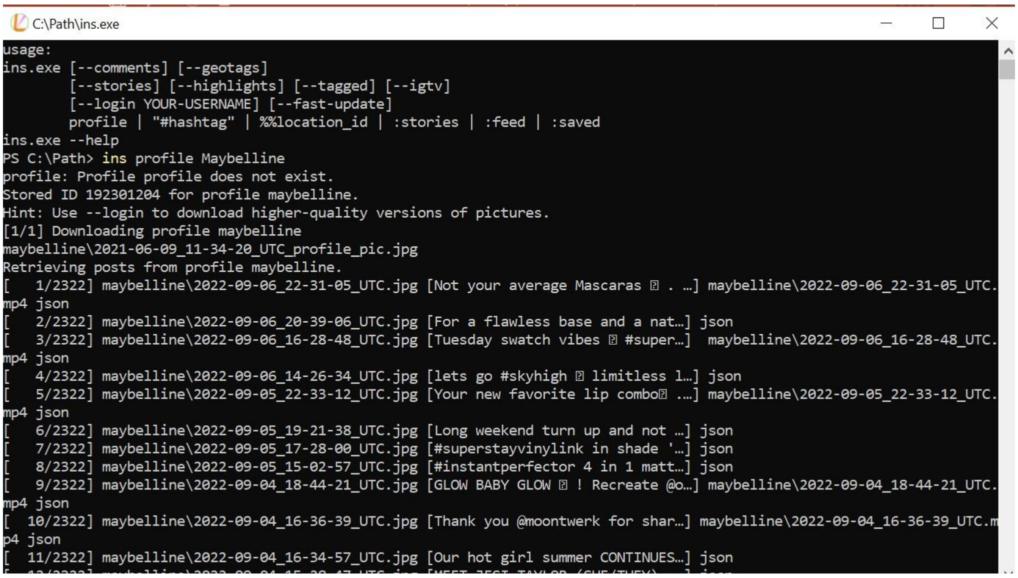


Fig 5.2: Instaloader Application - Data Scraping.

4.2 Tools

For this study, following tools were used.

1) Jupyter Notebook was used to carry out the development of machine learning model in python.
2) Instaloader.exe application for data scraping and Windows command prompt.

4.3 Data Pre-Processing

The data pre-processing step is crucial before developing Machine Learning based prediction model. The dataset was pre-processed as discussed earlier in the Section 3.1.

The images, json and text files for each brand is residing in their respective folder. The Images and Json forms the first part of the dataset which will be labelled and later merged. However, the textual dataset runs independently and though labelled or unlabelled, it won't have much impact as the sentiment analysis will be running brand wise.

4.3.1 Json and Image dataset Preprocessing

The strategy followed in labelling the dataset was by renaming each file prefixed with image_id (sequence number) and label (0: Loreal, 1: Lakme, 2: Maybelline).

Steps followed for labelling the dataset:

1) Access the directory path for each brand.
2) Search for the image file and rename it such that, if 2021_06_24_09_57_40_UTC.jpg is the name of the file then it must be renamed to 4_1_2021_06_24_09_57_40_UTC.jpg, where 1 denotes that it belongs to Lakme brand and 4 indicates that it is the 4th image of the brand Lakme.
3) Once renamed move the file to image folder of lakme.

4) Check if the json file exist for the same image file, if exist then rename the file similar to that we did for the image 2021_06_24_09_57_40_UTC.json to 4_1_2021_06_24_09_57_40_UTC.json and move it to the json folder of the brand Lakme, if it doesn't exist skip that file.

Below Fig 6.1 & 6.2 are the example of how the files will look like after being renamed.

Fig 6.1: Lakme Image File

Fig 6.2: Lakme Json File

Later the Json feature will be merged with the image feature using the image_id and label which form a unique primary key for them.

4.3.2 Json and Image features – Used for Popularity Prediction

- Image Features:

As discussed earlier, in addition to the deep learning feature extraction methods, there are a number of other methods that, when used in the appropriate combination, can improve the performance for popularity prediction. Additionally, the combination of the Gabor filter, the Local Binary Pattern (LBP), and the Gray Level Co-occurrence matrix (GLCM) has proven to be the best combination achieving accuracy of close 100%, as mentioned in the literature review section. We have considered the LBP entropy and LBP energy from the Local Binary Pattern since we would require a scalar number for data processing. Gray level co-occurrence matrix was used to examine contrast, homogeneity, dissimilarity, energy, and corelation characteristics, and Gabor Energy & Entropy evaluated using Gabor filter. Together, they made up the 9 features of the image. For each image that underwent feature extraction, an image id and label were also added.

	lbp_energy	lbp_entropy	contrast	dissimilarity	homogeneity	energy	correlation	gabor_energy	gabor_entropy	img_id	label
0	0.026346	0.956897	0.157181	0.379301	0.319580	0.045965	0.574594	0.774426	0.068853	1001	0
1	0.005590	0.989673	0.045160	0.152404	0.189090	0.015831	0.568563	0.813636	0.014531	1002	0
2	0.118544	0.802704	0.017742	0.085994	0.531246	0.261804	0.637545	0.403128	0.343829	1003	0
3	0.579147	0.353196	0.042068	0.070620	0.839335	0.625312	0.569915	0.800937	0.036334	1004	0
4	0.559808	0.374274	0.045253	0.077198	0.827927	0.613077	0.566243	0.798391	0.040364	1006	0
...	...	...	...	...	...	...	...	...	...	...	...
3430	0.006281	0.988263	0.022914	0.182470	0.144739	0.010517	0.627390	0.408598	0.392567	997	2
3431	0.010354	0.981935	0.096612	0.217840	0.268973	0.017882	0.569530	0.786708	0.053772	998	2
3432	0.100012	0.829068	0.003720	0.046249	0.562174	0.069001	0.648937	0.551779	0.241701	999	2
3433	0.065633	0.881328	0.080025	0.152023	0.413856	0.034136	0.604201	0.809540	0.021872	99	2
3434	0.363642	0.535669	0.009888	0.019452	0.862888	0.267230	0.626381	0.816409	0.009482	9	2

Fig 6.3: Image Features

- Json Features

By examining the three main categories of social features, we analyse the important role of context information connected with images; 1)Likes Count 2)Comments Count 3)Follows. Additionally, the dimension features does play an important role in popularity prediction and therefore, height and width has been account. Similar to images, label and img_id(=json_id) is being with json file, while feature extraction.

	image_id	comments_count	likes_count	height	width	category	follows	img_id	label
0	2871270766674398747	1	229	1333	750	Lakme	329.0	1005	1
1	2871300789512666062	4	701	1080	1080	Lakme	111.0	1006	1
2	2871331849608818788	1	317	1350	1080	Lakme	329.0	1007	1
3	2871334239070488163	3	256	1080	1080	Lakme	329.0	1008	1
4	2871394885240732850	2	5	1080	1080	Lakme	219.6	1009	1
...	...	...	...	...	...	...	...	...	...
2015	2638110990791485300	21	4951	750	750	Maybelline	1323.0	994	2
2016	2638201077108735455	42	4966	1333	750	Maybelline	1323.0	995	2
2017	2638774984673978612	101	16255	1080	1080	Maybelline	1323.0	997	2
2018	2638893320694026156	53	9424	1920	1080	Maybelline	1323.0	999	2
2019	1910697495903931291	19	609	1080	1080	Maybelline	5.0	9	2

2020 rows × 9 columns

Fig 6.4: Json features

4.3.3 Textual Dataset Pre-processing – Used for Sentiment Analysis

Moreover, in this research we carried and additional pre-processing with textual data for classification of the sentences. Data Pre-processing involves Remove punctions, emojis and URL, tokenisation and porter stemming.

4.3.3.1 Functions for textual Data Cleaning:

- Remove Url: For example, if there is a text in the file such as "Tweet checker http://ukdreams.com/files result". After calling a function remove_url, the resultant will be as below.

  Outcome will be the "Tweet checker result" and url http://ukdreams.com/files will be removed.

- Remove emojis: Remove all the emojis from the text. For example, if there a text such as "Smile face 😉".

  Outcome will be 'Smile face '.

- Remove Newline: Remove lines from the text. For example, if there is a text such as "Lakme has launched a product for hair which gives you silky shiny hair.
  Grows strength of the hair"

  Outcome will be "Lakme has launched a product for hair which gives you silky shiny hair. Grows strength of the hair."

- Remove Punctuations: Remove punctuation from the text. For example if the text is "`'!hi. wh?at is the cof[f]ee lik?e.'`

  Outcome will be "hi what is the coffee like".

Below figure Fig 7.1, represents the clean dataset for brand Lakme, after the above function have been applied, as it can be seen there are no punctuations, emojis, URL or newline(\n).

Similar kind of dataset has been formed for the other two brands dataset, L'Oreal and Maybelline.

	Text
0	Whos going to indulge in some self care tonight The LOrealMen purifying tissue face mask is the ultimate care for oily skin Enriched in purifying Oak Charcoal and hydrating glycerin it moisturiz...
1	Tired of looking tired Its tiring A little tip brought to you by jonathancohens to keep your skin healthy and glowing Fatigué davoir lair fatigué Cest fatiguant jonathancohens vous donne ses co...
2	Fight signs of fatigue on a daily basis thanks to HydraEnergetic Show your skin some love Combattez les signes de fatigue au quotidien grâce à HydraEnergetic Montrez à votre peau que vous laime...
3	They dont know each other but they share something the antifatigue HydraEnergetic Ils ne se connaissent pas mais ils ont un point commun le soin antifatigue HydraEnergetic jonathancohens can...
4	Get that energy boost Try our LOrealMen Healthy Look Express Care if youre looking for that perfect natural glow Its just 2 pumps away LorealMenExpert LorealMen MenExpert HealthyLook ExpressCare...
...	...
1117	STEAMPOD CARE Steampod 30 is all about the freedom to go from one look to another on all hair types To do so it is important to protect your hair before and after styling Steampod care routine ...
1118	SUMMER WAVES Hair by bencepalyi Going to a Summer party but no hair inspiration Here is bencepalyis hair mood inspiration for today Go for glam bouncy waves with Steampod 30 Dont forget to apply...
1119	STEAMPOD Does high shine ring a bell Then Steampod 30 doesnt need to be introduced Our professional steam styler transforms hair 2x faster¹ 2x smoother¹ and with 91 less damage² Dream it style ...
1120	LOOSE WAVES Hair by mane_ivy Did someone said boldbut natural Get your natural wavy Summer look with Steampod 30 whatever your hair type Nothing but Summer vibes powered by steam HairPros Which...
1121	STEAMPOD CREAM The joy of smoothing repairing Summer means sunbath salty hair and Steampod care with the Repairing smoothing cream that smoothes and tames thick unruly hair while protecting it...

Fig 7.1: Cleaned Textual data after the functions has been applied.

4.3.3.2 Language Detection and Translation:

Translation using translator function of googletrans module:

Ensure the text is in English by using the googletrans module's translator. The text should be kept if it is written in English. If not, identify the language and translate it into English language. In some cases, if the text is mixed and unable to translated or modified ignore the text and replace it with Null values.

Below Fig 7.2, represents the result set after performing the translator operation,

	Text	language	english
0	durgapuja durgapujainusa durgamaa saree sareelover sareedraping sareelovers sareeindia sareestyle sareeblouse makeup makeuplover makeupaddict makeupjunkie beautymakeup beautyjunkie beautyaddict be...	af	
1	Lakme To Natural CC Cream Color Bronze Its With Pure Aloe Vera SPF PA Real Price From Nykaa Hello Cuties Must Buy This Lakme to CC Cream Its Give Full Flawless Coverage Fragrance Is ...	en	Lakme To Natural CC Cream Color Bronze Its With Pure Aloe Vera SPF PA Real Price From Nykaa Hello Cuties Must Buy This Lakme to CC Cream Its Give Full Flawless Coverage Fragrance Is ...
2	Get your hands on your favourite product of Lakme at half of its actual amount only at The Beauty Shop or call us now Grab this once in a lifetime offer of off now Lakme TheBeautyShop LakmeGel L...	en	Get your hands on your favourite product of Lakme at half of its actual amount only at The Beauty Shop or call us now Grab this once in a lifetime offer of off now Lakme TheBeautyShop LakmeGel L...
3	niveaindia Nivea moisturizer swissbeautycosmetics Swiss beauty compact powder concealer nybae nybae contour blueheavencosmetics Blue heaven primer glamcosmetic glam highlighter glamcosmeti...	en	niveaindia Nivea moisturizer swissbeautycosmetics Swiss beauty compact powder concealer nybae nybae contour blueheavencosmetics Blue heaven primer glamcosmetic glam highlighter glamcosmeti...
4	Natural Makeup jahnavi_ teamperfectframes makeupbynishta makeupbyzainabfaiz hair southindianweddings southindianbride explorepage explore exploremore trending transition tuesday ...	en	Natural Makeup jahnavi_ teamperfectframes makeupbynishta makeupbyzainabfaiz hair southindianweddings southindianbride explorepage explore exploremore trending transition tuesday ...

Fig 7.2: Translated Textual Dataset where either the language has been converted into English or replaced it with NULL

4.3.3.3 Stopwords, Stemming and Tokenisation:

Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. In our approach we have used Porter Stemmer which is a simple algorithm for stemming English language words developed in 1968.

Tokenization on the other hand is the process of breaking up a long block of text into tokens. The three main categories of tokenization are word, character, and subword.

Stopwords: Stopwords are any words that do not significantly add to the meaning of a phrase. They can be safely ignored without affecting the sentence's meaning. These are some of the most popular short function words for some search engines, along with the, is, at, which, and on. Stop words might be problematic in this situation when looking for phrases that contain them, especially in names like "The Who" or "Take That."

Fig 7.3, represents the dataset of textual data where data in the token column is the resultant of the functions implemented such as **stopwards** – which as discussed removes the irrelevant words that does not add any meaning to the sentence or text like “who”, “is”, “the”; **stemming** – which replaces the words that means the same to its centric value like “Imagine”,”Imagination” and “Imaginative” will be replaced by “Imagin” and **tokenisation** – which split the sentence on the basis of delimeter “Multimodal Analysis using instagram” to “Multimodal” “Analysis” “using” “instagram” .

	Unnamed: 0	Text	language	english	token
0	0		en	Whos going to indulge in some self care tonight The LOrealMen purifying tissue face mask is the ultimate care for oily skin Enriched in purifying Oak Charcoal and hydrating glycerin it moisturiz...	[care, charcoal, enriched, face, fightoilyskin, glycerin, going, hydrating, indulge, lorealmen, lorealmenexpert, mask, mattifies, menexpert, minutes, moisturizes, oak, oily, purecharcoal, purifyin...
1	1		en	Tired of looking tired Its tiring A little tip brought to you by jonathancohens to keep your skin healthy and glowing Fatigué davoir lair fatigué Cest fatiguant jonathancohens vous donne ses co...	[brought, canalplus, canalplusseries, cest, conseils, davoir, donne, et, fatiguant, fatigué, glowing, healthy, hydraenergetic, jonathancohens, keep, laflamme, lair, little, looking, menexpert, men...
2	2		fr		[nan]
3	3		en	They dont know each other but they share something the antifatigue HydraEnergetic Ils ne se connaissent pas mais ils ont un point commun le soin antifatigue HydraEnergetic jonathancohens can...	[antifatigue, canalplus, canalplusseries, commun, connaissent, dont, hydraenergetic, ils, jonathancohens, know, laflamme, le, mais, menexpert, menskincare, ne, ont, pas, point, se, share, soin, so...
4	4		en	Get that energy boost Try our LOrealMen Healthy Look Express Care if youre looking for that perfect natural glow Its just pumps away LorealMenExpert LorealMen MenExpert HealthyLook ExpressCare ...	[away, boost, care, energy, express, expresscare, get, glow, healthy, healthylook, keephydrated, look, looking, lorealmen, lorealmenexpert, menexpert, natural, perfect, pumps, try, youre]
5	5		en	Dont let your energy fade LOrealMen Healthy Look Express Care is here to reenergize your skin LorealMenExpert MenExpert HealthyLook ExpressCare KeepHydrated	[care, dont, energy, express, expresscare, fade, healthy, healthylook, keephydrated, let, look, lorealmen, lorealmenexpert, menexpert, reenergize, skin]
6	6		en	Essentials to recharge Looking forward to good times and good laughs with all our friends Using the right products to energize your skin Were here to help you with the skin part Reenergize your...	[energetic, energize, essentials, forward, friends, good, help, hydra, hydraenergetic, laughs, line, look, looking, lorealman, lorealmenexpert, menexpert, moisturizer, part, products, putthatenerg...
7	7		en	Whos in need of some sunshine Weve got you covered Discover our Healthy Look Express Care an undetectable tinted gel powered with natural origin Vitamin CG that will leave skin feeling relaxed an...	[care, cg, covered, digital, discover, even, express, expresscare, feeling, filter, gel, get, glow, got, h, healthy, healthylook, keephydrated, leave, look, lorealmen, lorealmenexpert, menexpert, ...
8	8		en	Time to wake your skin up Our LOrealMen Hydra Energetic line is here to boost your skin pump it with vitamin CG to leave you feeling refreshed and ready throughout your day LorealMenExpert Lorea...	[boost, cg, day, energetic, feeling, hydra, hydraenergetic, leave, line, lorealmen, lorealmenexpert, menexpert, moisturizer, pump, putthatenergyback, ready, refreshed, skin, throughout, time, vita...
9	9		en	Name a more iconic duo Our icecool eye rollon and Healthy Look Express care make the perfect pair to fight dark circles and bring you that summer glow any time of the year LorealMenExpert Loreal...	[bring, care, circles, dark, duo, express, expresscare, eye, fight, glow, healthy, hydraenergetic, icecool, iconic, keephydrated, look, lorealmen, lorealmenexpert, make, menexpert, moisturizer, na...

Fig 7.3: Resultant Dataset after removing the stopwords, applied stemming and tokenising the word

4.4: Methodology – Modelling and Evaluation of Results

In this section we will discuss the machine learning models and the NLP pretrained model used for the popularity prediction and textual sentiment analysis respectively. The section 4.4.1 discusses the data splitting.

4.4.1: Data Splitting for Training and Testing

The combined dataset of JSON and image features was used for model optimization. Before the dataset was combined, it was split into training and testing. The dataset was divided into two halves using the criteria of 80% training data and 20% testing data. Module splitfolder has been used to carry out the splitting operation.

4.4.2: Popularity prediction using XGBoost Dmatrix and XGBoost Regressor.

It is an internal data structure used by XGBoost which is optimized for memory efficiency and training speed. It has an inbuilt cross validation feature.

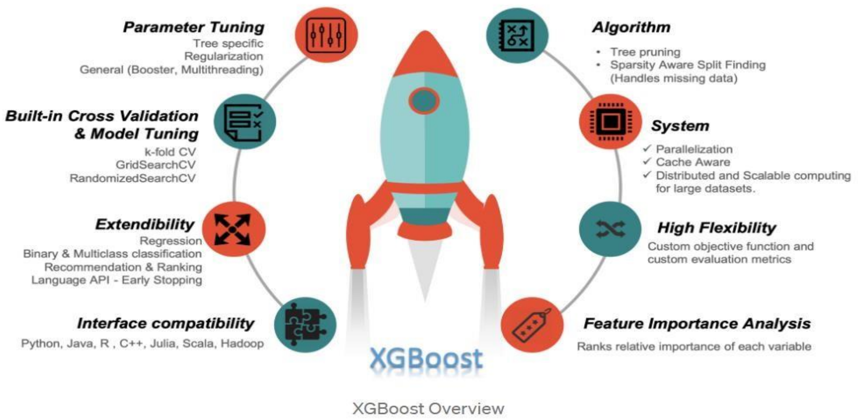


Fig 8: XGBoost Overview (Ref: https://medium.com/sfu-cspmp/xgboost-a-deep-dive-into-boosting-f06c9c41349)

As discussed earlier, XGBoost has a bunch of benefits such as Tree pruning, parallelization, high flexibility, extendibility to regression, classification and Language API and also parameter tuning which includes regularization, boosting and multithreading.

With our approach, we have implemented the XGBoost DMatrix and XGBoost regression to predict the popularity using number of likes as our target feature along with its respective label (0: Loreal, 1: Lakme, 2: Maybelline) and training the model on the dataset with the best hyperparameters selection to enhance the predictability power of our popularity prediction model.

4.4.2.1 XGBoost Dmatrix Model.

After multiple tried and test methods below were the chosen value for XGBoost DMatrix parameters. Though we have taken multiple attempts to decide on the parameters, we have kept the top 3 best model and its respective parameters in the code (one for DMatrix and two for XGBoost Regression).

The parameters best chosen for DMatrix were as follows:

max_depth (Best Range: 3 to 10): Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit. In our approach we have taken value as 7.

eta (Learning rate: Default = 0.6): After each boosting step, we can directly get the weights of new features, and eta shrinks the feature weights to make the boosting process more conservative. In our approach we have taken the value as 0.2.

objective [default=reg:squarederror]: Considered same as default - reg:squarederror: regression with squared loss.

nthread (Defaults to maximum number of threads if not set): Number of parallel threads used to run XGBoost. In our approach, considering the system capability taken a value as 5.

eval_metric: Evaluation metrics for validation data. As we are dealing with the regression model we have considered RMSE as the value.num_round: The number of rounds for boosting. We have taken it as 50, as the value tends to get stable post that.

When evaluated with these features our XGBoost DMatrix gets us the best result with respect to root mean squared error (RMSE) values.

```
# Train the model.
num_round = 50
bst = xgb.train(param, dtrain, num_round, evallist)
```

```
C:\Users\Sharoz\anaconda3\lib\site-packages\xgboost\core.py:525: FutureWarning: Pass `evals` as keyword args.  Passing these as
positional arguments will be considered as error in future releases.
  warnings.warn(
```

```
[0]	eval-rmse:0.58543	train-rmse:0.58144
[1]	eval-rmse:0.47276	train-rmse:0.46694
[2]	eval-rmse:0.37949	train-rmse:0.37431
[3]	eval-rmse:0.30499	train-rmse:0.30016
[4]	eval-rmse:0.24564	train-rmse:0.24077
[5]	eval-rmse:0.19825	train-rmse:0.19321
[6]	eval-rmse:0.16060	train-rmse:0.15520
[7]	eval-rmse:0.13050	train-rmse:0.12474
[8]	eval-rmse:0.10713	train-rmse:0.10031
[9]	eval-rmse:0.08892	train-rmse:0.08079
[10]	eval-rmse:0.07489	train-rmse:0.06520
[11]	eval-rmse:0.06403	train-rmse:0.05269
[12]	eval-rmse:0.05667	train-rmse:0.04294
[13]	eval-rmse:0.05088	train-rmse:0.03503
[14]	eval-rmse:0.04729	train-rmse:0.02892
[15]	eval-rmse:0.04453	train-rmse:0.02392
[16]	eval-rmse:0.04312	train-rmse:0.01994
[17]	eval-rmse:0.04171	train-rmse:0.01698
[18]	eval-rmse:0.04100	train-rmse:0.01470
[19]	eval-rmse:0.04037	train-rmse:0.01307
[20]	eval-rmse:0.04018	train-rmse:0.01177
[21]	eval-rmse:0.04003	train-rmse:0.01091
[22]	eval-rmse:0.03986	train-rmse:0.01017
[23]	eval-rmse:0.03974	train-rmse:0.00969
[24]	eval-rmse:0.03989	train-rmse:0.00910
[25]	eval-rmse:0.03996	train-rmse:0.00870
[26]	eval-rmse:0.03992	train-rmse:0.00847
[27]	eval-rmse:0.03995	train-rmse:0.00818
[28]	eval-rmse:0.03991	train-rmse:0.00795
[29]	eval-rmse:0.03980	train-rmse:0.00774
[30]	eval-rmse:0.03980	train-rmse:0.00766
[31]	eval-rmse:0.03979	train-rmse:0.00753
[32]	eval-rmse:0.03977	train-rmse:0.00749
[33]	eval-rmse:0.03978	train-rmse:0.00742
[34]	eval-rmse:0.03976	train-rmse:0.00719
[35]	eval-rmse:0.03976	train-rmse:0.00705
[36]	eval-rmse:0.03971	train-rmse:0.00685
[37]	eval-rmse:0.03965	train-rmse:0.00670
[38]	eval-rmse:0.03963	train-rmse:0.00663
[39]	eval-rmse:0.03964	train-rmse:0.00652
[40]	eval-rmse:0.03957	train-rmse:0.00633
[41]	eval-rmse:0.03957	train-rmse:0.00631
[42]	eval-rmse:0.03956	train-rmse:0.00607
[43]	eval-rmse:0.03959	train-rmse:0.00591
[44]	eval-rmse:0.03959	train-rmse:0.00564
[45]	eval-rmse:0.03961	train-rmse:0.00552
[46]	eval-rmse:0.03960	train-rmse:0.00550
[47]	eval-rmse:0.03967	train-rmse:0.00528
[48]	eval-rmse:0.03966	train-rmse:0.00514
[49]	eval-rmse:0.03973	train-rmse:0.00510
```

Fig 9.1: Evaludation of XGBoost Dmatrix with num_rounds taken as 50.

The above Fig 9 represents the training of dataset and simultaneously calculating the rmse on both training as well as on the validation dataset. Train-rmse:0.00510 is the accuracy we received with our training dataset which is exceptional. Although, the root mean square value on the eval or validation dataset is not as good as on training dataset but still is extraordinary

Following this, the model has been trained and can be used for prediction of target value which is combination of number of likes and label.

The RMSE values calculated on the predicted values is 0.039728, which to the best of my knowledge beats the previous studies in the area of multimodal popularity prediction.

As it can be seen from the below Fig 9.1, the predicted values for label matched the predicted and the actual values and even predicted number of likes seems to be very close to actual values.

	Predlikes	Predlabel	Actuallikes	Actuallabel
0	0.008097	0.000105	0.012539	0
1	0.002987	-0.000093	0.001912	0
2	0.023363	0.000301	0.033895	0
3	0.022239	-0.000015	0.033702	0
4	0.029992	0.000105	0.007932	0
...	...	...	...	...
101	0.138502	1.999945	0.102641	2
102	0.089447	1.999945	0.115772	2
103	0.002155	1.999929	0.002011	2
104	0.087720	1.999945	0.098874	2
105	0.212060	1.999945	0.161479	2

Fig 9.2: Comparison of Actual and Predicted values for label and number of likes

4.4.2.2 XGBoost Regressor.

As discussed, we have a multioutput – number of likes and label; Hence to apply any regression technique on the model, we have to utilize Multioutput regressor on the top of the selected model and the set of parameters to be select for training of model has to be of multioutput regressor which is in correlation with any regression model selected.

The model is implemented with GridSearchCV for best hyperparameter selection. Although, conducted multiple attempts to select the best hyperparameter, we have shown below two sets of hyperparameter on which the result was exceptional.

- Parameters – 1st Set estimator__min_child_weight: It is the minimum sum of instance weight needed in a child. If the tree partition step results in a leaf node with the sum of instance weight less than min_child_weight, then the building process will give up further partitioning.In our approach we have considered it to be 4.

estimator__learning_rate: After each boosting step, we can directly get the weights of new features, and eta shrinks the feature weights to make the boosting process more conservative. In our approach we have taken the list of values [.03, 0.05, 0.1] estimator__max_depth: For max_depth as the ideal range is in between 3 to 10, we have tried training our model with values as [4, 5, 6]

estimator__n_estimators: It refers to number of trees the model will be forming. For our model we have considered a list of values [500, 700, 1000].

estimator__eval_metric: We have chosen RMSE as the evaluation for all the regression model, so as to build a comparison between at a later stage.

```
GridSearchCV(cv=2,
             estimator=MultiOutputRegressor(estimator=XGBRegressor(base_score=None,
                                                                   booster=None,
                                                                   callbacks=None,
                                                                   colsample_bylevel=None,
                                                                   colsample_bynode=None,
                                                                   colsample_bytree=None,
                                                                   early_stopping_rounds=None,
                                                                   enable_categorical=False,
                                                                   eval_metric=None,
                                                                   gamma=None,
                                                                   gpu_id=None,
                                                                   grow_policy=None,
                                                                   importance_type=None,
                                                                   interaction_constraints=None,
                                                                   learning...
                                                                   monotone_constraints=None,
                                                                   n_estimators=100,
                                                                   n_jobs=None,
                                                                   num_parallel_tree=None,
                                                                   predictor=None,
                                                                   random_state=None,
                                                                   reg_alpha=None,
                                                                   reg_lambda=None, ...)),
             n_jobs=5,
             param_grid={'estimator__eval_metric': ['rmse'],
                         'estimator__learning_rate': [0.03, 0.05, 0.07],
                         'estimator__max_depth': [5, 6, 7],
                         'estimator__min_child_weight': [4],
                         'estimator__n_estimators': [500, 700, 1000]},
             verbose=True)
```

Fig 10.1: XGBOOST regressor implemented with GridSearchCV with defined parameters (1st Set).

The RMSE for the selected hyperparameter values and predicted on validation dataset comes out to be 0.039527, which is close to RMSE of XGBoost DMatrix and again is extraordinary in this field.

Hence, XGBoost DMatrix and XGBoost Regressor has both proven to be the best performing model for this approach with an exceptional RMSE value of 0.039728 and 0.39527 respectively.

- Parameters – 2nd Set estimator__min_child_weight: It is the minimum sum of instance weight needed in a child. If the tree partition step results in a leaf node with the sum of instance weight less than min_child_weight, then the building process will give up further partitioning. In our approach we have considered it to be [10,20].

estimator__learning_rate: After each boosting step, we can directly get the weights of new features, and eta shrinks the feature weights to make the boosting process more conservative. In our approach we have taken the list of values [.01, 0.08, 0.2].

estimator__max_depth: For max_depth as the ideal range is in between 3 to 10, we have tried training our model with values as [3,5,9]

estimator__n_estimators: It refers to number of trees the model will be forming. For our model we have considered a list of values [200, 300, 700]

estimator__eval_metric: We have chosen RMSE as the evaluation for all the regression model, so as to build a comparison between at a later stage.

```
: GridSearchCV(cv=5,
             estimator=MultiOutputRegressor(estimator=XGBRegressor(base_score=None,
                                                                    booster=None,
                                                                    callbacks=None,
                                                                    colsample_bylevel=None,
                                                                    colsample_bynode=None,
                                                                    colsample_bytree=None,
                                                                    early_stopping_rounds=None,
                                                                    enable_categorical=False,
                                                                    eval_metric=None,
                                                                    gamma=None,
                                                                    gpu_id=None,
                                                                    grow_policy=None,
                                                                    importance_type=None,
                                                                    interaction_constraints=None,
                                                                    learning...
                                                                    monotone_constraints=None,
                                                                    n_estimators=100,
                                                                    n_jobs=None,
                                                                    num_parallel_tree=None,
                                                                    predictor=None,
                                                                    random_state=None,
                                                                    reg_alpha=None,
                                                                    reg_lambda=None, ...)),
             n_jobs=5,
             param_grid={'estimator__eval_metric': ['rmse'],
                         'estimator__learning_rate': [0.01, 0.08, 0.2],
                         'estimator__max_depth': [3, 5, 9],
                         'estimator__min_child_weight': [10, 20],
                         'estimator__n_estimators': [200, 300, 700]},
             verbose=True)
```

Fig 10.2: XGBOOST regressor implemented with GridSearchCV with defined parameters (2nd Set).

The RMSE for the selected hyperparameter values and predicted on validation dataset comes out to be 0.077632, which is lower than the previous XGBoost DMatrix model and XGBoost Regressor trained on 1st Set of hyperparameter.

4.4.3: Random Forest Regression.

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their average for regression.

As mentioned earlier, we have a multioutput – number of likes and label; Hence to apply any regression technique on the model, we have to utilize Multioutput regressor on the top of the selected model

The model is implemented with GridSearchCV for best hyperparameter selection.

- HyperParameters:

estimator_max_depth: It is the maximum depth of a tree, use to control the overfitting as higher depth will allow model to learn relation very specific to particular sample. For our approach the values chosen to be are [2,3,5,10,20].

estimator_min_sample_leaf (int or float, Default =1): The minimum number of samples required to be at a leaf node. For our approach the values chosen to be [5,10,20,50,100,200].

estimator_n_estimators (int, Default = 100): the number of trees in the forest. For our approach the values chosen to be [10,25,30,50,100,200].

```
GridSearchCV(cv=4,
             estimator=MultiOutputRegressor(estimator=RandomForestRegressor(random_state=42)),
             n_jobs=-1,
             param_grid={'estimator__max_depth': [2, 3, 5, 10, 20],
                         'estimator__min_samples_leaf': [5, 10, 20, 50, 100,
                                                         200],
                         'estimator__n_estimators': [10, 25, 30, 50, 100, 200]},
             verbose=1)
```

Fig 11: Random Forest regression implemented with GridSearchCV with defined parameters.

The RMSE for the selected hyperparameter values and predicted on validation dataset comes out to be 0.3294, which is far lower than XGBoost Dmatrix and XgBoost Regressor, hence not that efficient and effective for our Dataset.

4.4.4: SVR and Linear Regression

SVR and Linear Regression – Dataset has been trained on default hyperparameters of the models, except for one parameter being defined while training the SVR which is epsilon;It specifies the epsilon-tube within which no penalty is associated in the training loss function with points predicted within a distance epsilon from the actual value (the area within the decision boundary). Again, a note that multioutput regressor has been performed on top of both the model to achieve the objective. Both the models have proven to be effective but fails to impress when compared with XGBoost Dmatrix and XGBoost Regressor. Although, Linear Regression seems to be in close precision in terms of RMSE value as 0.32705, with random forest regressor which is impressive.

4.5: Sentiment Analysis using Text Blob

Following the section 4.3.3, the textual dataset was cleaned to carry the process of sentimental analysis of the caption extracted for each brand - Lakme, L'Oreal and Maybelline.

The pretrained NLP model TextBlob was used to carry out the sentiment analysis. TextBlob returns two parameters such as subjectivity and polarity. For our approach we have only considered the Polarity as we were focussed in the defining whether the text was positive, negative or neutral. The range of polarity for Textblob as discussed earlier is [-1,1], where -1 indicates negative polarity of text and 1 indicates positive polarity, whereas 0 indicates neutral.

Table 2, Table 3, Table 4, represents the polarity score calculated on textual dataset of each brand – Lakme, L'Oreal and Maybelline using TextBlob model.

language	english	token	sentiment
en	Whos going to indulge in some self care tonight The LOrealMen purifying tissue face mask is the ultimate care for oily skin Enriched in purifying Oak Charcoal and hydrating glycerin it moisturiz...	[care, charcoal, enriched, face, fightoilyskin, glycerin, going, hydrating, indulge, lorealmen, lorealmenexpert, mask, mattifies, menexpert, minutes, moisturizes, oak, oily, purecharcoal, purifyin...	0.0
en	Tired of looking tired Its tiring A little tip brought to you by jonathancohens to keep your skin healthy and glowing Fatigué davoir lair fatigué Cest fatiguant jonathancohens vous donne ses co...	[brought, canalplus, canalplusseries, cest, conseils, davoir, donne, et, fatiguant, fatigué, glowing, healthy, hydraenergetic, jonathancohens, keep, laflamme, lair, little, looking, menexpert, men...	-0.121875
fr		[nan]	0.0
en	They dont know each other but they share something the antifatigue HydraEnergetic Ils ne se connaissent pas mais ils ont un point commun le soin antifatigue HydraEnergetic jonathancohens can...	[antifatigue, canalplus, canalplusseries, commun, connaissent, dont, hydraenergetic, ils, jonathancohens, know, laflamme, le, mais, menexpert, menskincare, ne, ont, pas, point, se, share, soin, so...	-0.125
en	Get that energy boost Try our LOrealMen Healthy Look Express Care if youre looking for that perfect natural glow Its just pumps away LorealMenExpert LorealMen MenExpert HealthyLook ExpressCare ...	[away, boost, care, energy, express, expresscare, get, glow, healthy, healthylook, keephydrated, look, looking, lorealmen, lorealmenexpert, menexpert, natural, perfect, pumps, try, youre]	0.533333

Table 2: Sentiment Score (Polarity) for L'Oreal Brand

language	english	token	sentiment
en	Bold Crimson a classic matte red that suits all	[bold, classic, crimson, matte, red, suits]	0.166667
en	Maybelline Creamy Mattes in Lively Violet a muted cool pink thats universally flattering especially on morenas	[cool, creamy, especially, flattering, lively, mattes, maybelline, morenas, muted, pink, thats, universally, violet]	0.077273
en	Barely Nude a light peachy nude	[barely, light, nude, peachy]	0.225
en	Say goodbye to visible eye bags with Instant Age Rewind Undeniably the best eye bag concealer	[age, bag, bags, best, concealer, eye, goodbye, instant, rewind, say, undeniably, visible]	0.5
en	Who wouldnt want these lashes Shop Hypercurl now	[hypercurl, lashes, shop, want, wouldnt]	0.0
en	This is not a drill The USA bestseller mascara with allaround star reviews is coming in DAYS Isnt it TOTALLY TEMPTING Get yours first on October exclusive on Shopee TotalTemptationPH	[allaround, bestseller, coming, days, drill, exclusive, first, get, isnt, mascara, october, reviews, shopee, star, tempting, totally, totaltemptationph, usa]	0.125
en	DAYS TO GO before you can get your very own Total Temptation mascara With its CREAMY WHIPPED FORMULA youre sure to have the SOFTEST LASHES ever Totally awesome right TotalTemptationPH	[awesome, creamy, days, ever, formula, get, go, lashes, mascara, right, softest, sure, temptation, total, totally, totaltemptationph, whipped, youre]	0.513143
en	One more reason to LOOOOVE Total Temptation it smells like coconut and is sure to give you TOTALLY SOFT lashes DAYS TO GO LADIES Are you ready Get yours exclusively on Shopee on October TotalTem...	[coconut, days, exclusively, get, give, go, ladies, lashes, like, loooove, october, one, ready, reason, shopee, smells, soft, sure, temptation, total, totally, totaltemptationph]	0.26
en	Check out the length of those lashes DAYS TO GO GIRLS Get ready to buy your Total Temptation mascara on October am exclusive on Shopee Limited stocks only TotalTemptationPH	[buy, check, days, exclusive, get, girls, go, lashes, length, limited, mascara, october, ready, shopee, stocks, temptation, total, totaltemptationph]	0.032143
en	Marvel at these gorgeous shades from our Superstay Matte Ink collection thats perfect for any skin tone MaybellinePH	[collection, gorgeous, ink, marvel, matte, maybellineph, perfect, shades, skin, superstay, thats, tone]	0.85

Table 3: Sentiment Score (Polarity) for Maybelline Brand

language	english	token	sentiment
af		[care, charcoal, enriched, face, fightoilyskin, glycerin, going, hydrating, indulge, lorealmen, lorealmenexpert, mask, mattifies, menexpert, minutes, moisturizes, oak, oily, purecharcoal, purifyin...	0.0
en	Lakme To Natural CC Cream Color Bronze Its With Pure Aloe Vera SPF PA Real Price From Nykaa Hello Cuties Must Buy This Lakme to CC Cream Its Give Full Flawless Coverage Fragrance Is ...	[brought, canalplus, canalplusseries, cest, conseils, davoir, donne, et, fatiguant, fatigué, glowing, healthy, hydraenergetic, jonathancohens, keep, laflamme, lair, little, looking, menexpert, men...	0.399429
en	Get your hands on your favourite product of Lakme at half of its actual amount only at The Beauty Shop or call us now Grab this once in a lifetime offer of off now Lakme TheBeautyShop LakmeGel L...	[]	-0.055556
en	niveaindia Nivea moisturizer swissbeautycosmetics Swiss beauty compact powder concealer nybae nybae contour blueheavencosmetics Blue heaven primer glamcosmetic glam highlighter glamcosmeti...	[antifatigue, canalplus, canalplusseries, commun, connaissent, dont, hydraenergetic, ils, jonathancohens, know, laflamme, le, mais, menexpert, menskincare, ne, ont, pas, point, se, share, soin, so...	0.147273
en	Natural Makeup jahnavi_ teamperfectframes makeupbynishta makeupbyzainabfaiz hair southindianweddings southindianbride explorepage explore exploremore trending transition tuesday ...	[away, boost, care, energy, express, expresscare, get, glow, healthy, healthylook, keephydrated, look, looking, lorealmen, lorealmenexpert, menexpert, natural, perfect, pumps, try, youre]	0.1

Table 4: Sentiment Score (Polarity) for Lakme Brand

The pretrained NLP model Vader was also used to carry out the sentiment analysis. Instead of making an effort to distinguish between a sentence's subjectivity, objectivity, or truthfulness. it is simply interested in whether the text communicates a positive, negative, or neutral opinion. In our approach we considered the compound value for our analysis.

Table 5, Table 6, Table 7, represents the compound polarity score calculated on textual dataset of each brand – Lakme, L'Oreal and Maybelline using Vader Sentiment analyser.

Text	language	english	token	sentiment
	af		[nan]	0.0
	en	Lakme To Natural CC Cream Color Bronze Its With Pure Aloe Vera SPF PA Real Price From Nykaa Hello Cuties Must Buy This Lakme to CC Cream Its Give Full Flawless Coverage Fragrance Is ...	[affordable, aloe, amazon, beauty, beautyblogger, bio, bronze, buy, cc, channel, color, coverage, cream, cuties, daily, elite_makeup_junkie, flawless, flipkart, folks, fragrance, full, getitwithlu...	0.9617
	en	Get your hands on your favourite product of Lakme at half of its actual amount only at The Beauty Shop or call us now Grab this once in a lifetime offer of off now Lakme TheBeautyShop LakmeGel L...	[actual, amount, beauty, call, discount, favourite, get, grab, half, hands, lakme, lakmecosmetics, lakmegel, lakmeindia, lakmeproducts, lifetime, makeup, makeupproducts, offer, product, shop, skin...	0.5859
	en	niveaindia Nivea moisturizer swissbeautycosmetics Swiss beauty compact powder concealer nybae nybae contour blueheavencosmetics Blue heaven primer glamcosmetic glam highlighter glamcosmeti...	[bb, beauty, blue, blueheavencosmetics, compact, concealer, contour, cream, definer, eyebrows, glam, glamcosmetic, good, good_vibesin, heaven, highlighter, lakme, lakmeindia, lipstick, moisturizer...	0.9231

Table 5: Sentiment Score (Vader compound) for Lakme Brand

Text	language	english	token	sentiment
	en	Bold Crimson a classic matte red that suits all	[bold, classic, crimson, matte, red, suits]	0.3818
	en	Maybelline Creamy Mattes in Lively Violet a muted cool pink thats universally flattering especially on morenas	[cool, creamy, especially, flattering, lively, mattes, maybelline, morenas, muted, pink, thats, universally, violet]	0.7579
	en	Barely Nude a light peachy nude	[barely, light, nude, peachy]	0.0
	en	Say goodbye to visible eye bags with Instant Age Rewind Undeniably the best eye bag concealer	[age, bag, bags, best, concealer, eye, goodbye, instant, rewind, say, undeniably, visible]	0.6369

Table 6: Sentiment Score (Vader compound) for Lakme Brand

Text	language	english	token	sentiment
	en	Whos going to indulge in some self care tonight The LOrealMen purifying tissue face mask is the ultimate care for oily skin Enriched in purifying Oak Charcoal and hydrating glycerin it moisturiz...	[care, charcoal, enriched, face, fightoilyskin, glycerin, going, hydrating, indulge, lorealmen, lorealmenexpert, mask, mattifies, menexpert, minutes, moisturizes, oak, oily, purecharcoal, purifyin...	0.7506
	en	Tired of looking tired Its tiring A little tip brought to you by jonathancohens to keep your skin healthy and glowing Fatigué davoir lair fatigué Cest fatiguant jonathancohens vous donne ses co...	[brought, canalplus, canalplusseries, cest, conseils, davoir, donne, et, fatiguant, fatigué, glowing, healthy, hydraenergetic, jonathancohens, keep, laflamme, lair, little, looking, menexpert, men...	-0.4767
	fr		[nan]	0.0

Table 7: Sentiment Score (Vader compound) for Lakme Brand

CHAPTER V: Results and Conclusions

Introduction

To demonstrate the effectiveness of the proposed multi-label multimodal machine learning XGBoost model, it is compared with several baselines. In section 5.1, this chapter discusses the prediction achieved by each machine learning model and compared them on the basis of the metric defined, which is root mean squared error (RMSE). In Addition, section 5.2, takes into account the most effective method and utilize it build the final conclusion using the predicted value of which brand gained the highest popularity out of three brands – L'Oreal, Lakme and Maybelline. Section 5.3 discusses an extra research where we have proposed a sentiment analysis on textual data and based on the results achieved compared the mean sentiment for each brand. Later, we build another conclusion of whether the textual sentimental result will have any impact on the improvement of popularity prediction or not.

5.1 Machine Learning models comparison based on Performance metric.

We compare the effectiveness of the machine learning models mentioned in the previous section using the performance metric root mean square error (RMSE).

Table 2 shows that, due to its outstanding RMSE value of 0.03952, the XGBoost Regressor significantly outperformed all previous baseline machine learning models (Random Forest Regressor, Support Vector Regressor and Linear Regressor) used in the field of multimodal popularity prediction. Additionally, given that we constructed three sets using the XGBoost regressor—XGBoost DMatrix, XGBoost Regressor with the First Set of Hyperparameters, and XGBoost Regressor with the Second Set of Hyperparameters—we can conclude that the XGBoost regressor with the First Set and the XGBoost Dmatrix techniques perform the best in terms of predicting the popularity of the multi-label multimodal popularity prediction, with a RMSE score of 0.03952 and 0.03972 respectively.

MODELS	RMSE
XGBoost Dmatrix	0.03972
XGBoost Regressor (1^{st} Set)	0.03952
XGBoost Regressor (2^{nd} Set)	0.07763
Random Forest Regressor	0.32945
Linear Regression	0.32705
Support Vector Regressor	0.36452

Table 8: Model Performance based on RMSE metric.

Even though XGBoost defeated other machine learning models including Random Forest Regressor, Support Vector Regressor, and Linear Regressor, they still have respectable RMSE scores. Additionally, these models' RMSEs are fairly competitive with one another, with Linear Regression and Randome Forest coming in at 0.32705 and 0.32945, respectively.

5.2 Comparison of Brands popularity using XGBoost

The XGBoost Regressor and XGBoost Dmatrix turn out to be the most successful model in forecasting the popularity prediction of multi-label multimodal, as stated in section 5.1. XGBoost DMatrix has been used to do this study so that we can compare the three brands side by side and draw conclusions on which one ends up being the most popular.

Fig 12.1 represents the distribution of likes inclusive of all the three brands. The number of likes has been taken as the X-axis, and frequency (number of images) has been taken as the yaxis. The number of likes ranges from 0 to 60,000. As from the graph, it can be seen very few images were able to achieve the highest mark and mostly were ranged between 0 to 10,000.

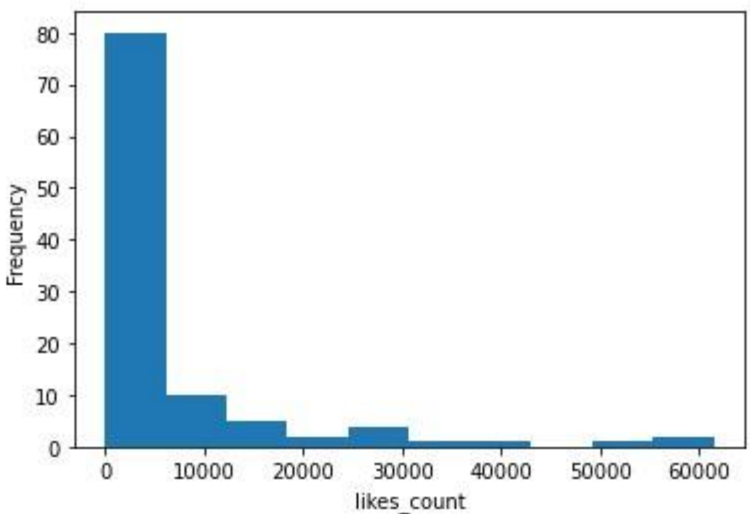


Fig 12.1: Distribution of likes inclusive of all three brands.

Similar to above, distribution of likes for the images were plotted for all the three brands to get an insight of the pattern followed by each of them.

Fig 12.2 represents the likes distribution of L'Oreal brand, it can be seen the maximum number the images have received are in the range of 25k, whereas most number of images lies within 5k limit of number of likes.

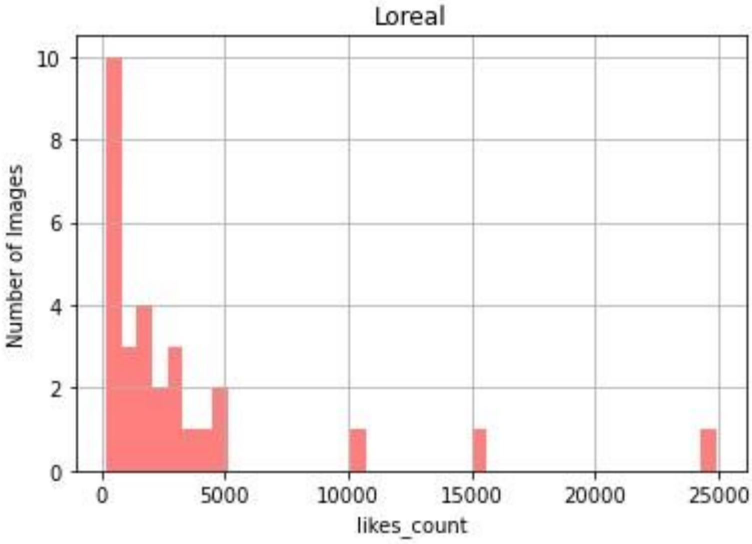


Fig 12.2: Likes distribution for Loreal Brand

Fig 12.3 represents the number of likes of Lakme brand. The maximum number of likes attained by Lakme brand is far less than that of L'Oreal and Maybelline and the most of images lies within the range of 2k number of likes. There were some images which recorded the 8k likes.

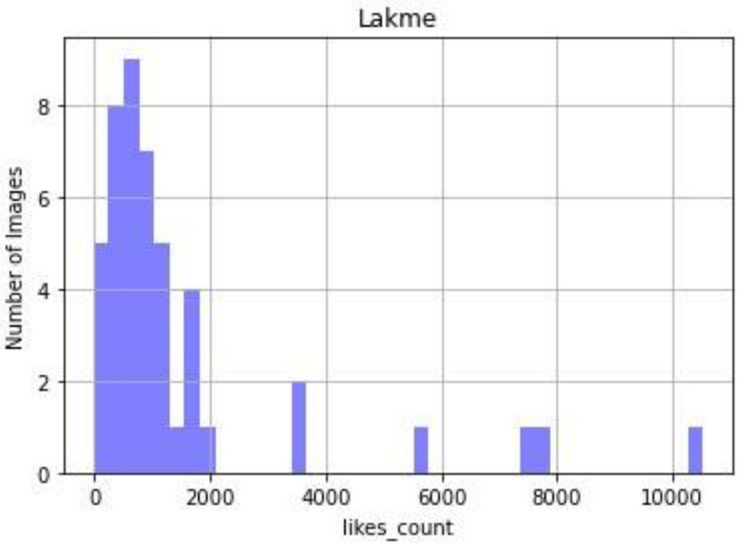


Fig 12.3: Likes distribution for Lakme Brand

Fig 12.4 represents the number of likes of Maybelline brand. Maybelline has shown better distribution of likes compared to the other two brands. The maximum number of likes attained by couple of images was 60k approximately, while few images were seen to be falling in the range of 20k to 40k number of likes. Whereas there were still quite a number of images which falls in the minimum range of 10k number of likes.

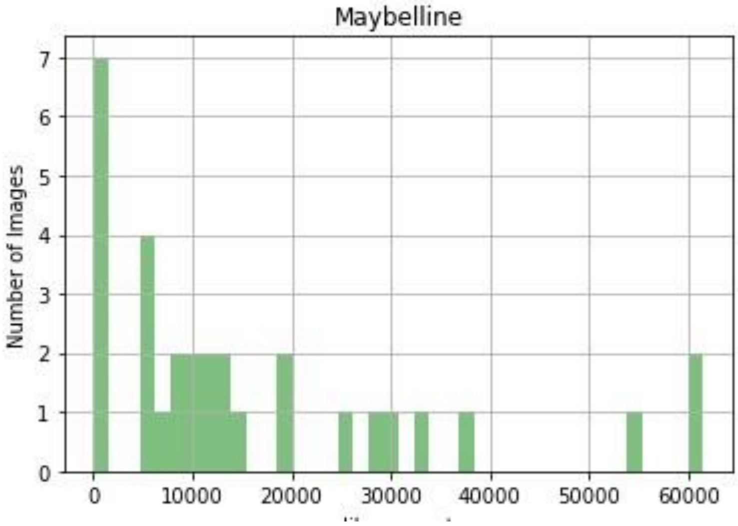


Fig 12.4: Likes distribution for Maybelline Brand

For Final analysis, we have considered the top 25 images from each brand and then taken the mean of number of likes for each of them. As it can be inferred from the distribution of likes as well as from the Fig 13, that the forecasted or predicted value of number of likes for Maybelline far exceeded its popularity from the other two brands Lakme and L'Oreal. Moreover, L'Oreal has acquired the second places whereas Lakme stood last out of three in terms of number of likes which is in direct corelation with the popularity.

This forms our final conclusion of the popularity prediction of the three stated brands.

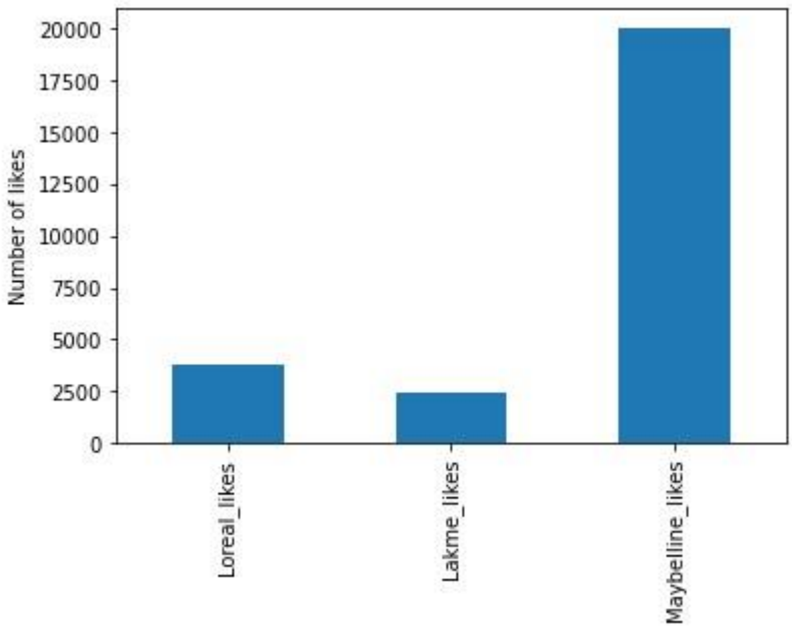


Fig 13: Mean of number of likes for top 25 images of all the three brands

5.3 Sentiment Analysis of Textual data and corelation with popularity prediction.

As discussed in the section 4.3.3, textual data was cleaned and pre-processed to carry out the sentiment analysis. TextBlob and Vader pretrained model has been used to calculate the sentiment score for each of the brand.

Fig 14 represents the mean of the sentiment score achieved on the textual dataset of each brand. The highest positivity was attained by Lakme brand of 0.25, whereas Maybelline also had the score close to that of Lakme of approximately 0.25. The mean of sentiment score for L'Oreal was about 0.15.

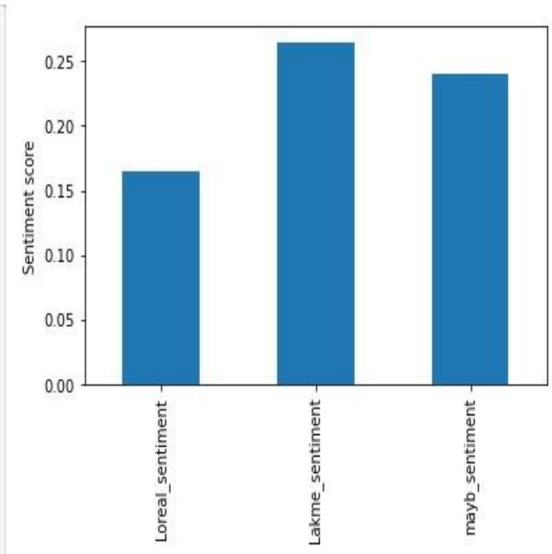


Fig 14.1: Mean of sentiment score for all three brands using TextBlob

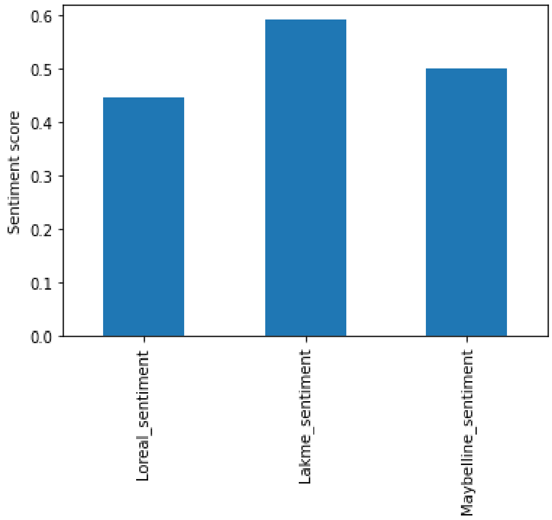


Fig 14.2: Mean of sentiment score for all three brands using Vader.

Fig 14.2 represents the mean of the sentiment score achieved on the textual dataset of each brand using vader sentiment analyser. The highest positivity was attained by Lakme brand of close to 0.6, whereas Maybelline also had the score close to that of Lakme of approximately 0.5. The mean of sentiment score for L'Oreal was about 0.45. Therefore, we can conclude that similar kind of pattern has been achieved with Vader model with a different sentiment score but all the brands lying approximately in the same range of sentiment score

Since no brand would reasonably promote their brand with negative emotion that could harm their popularity, it can be inferred from the values displayed above that the caption textual data for all three brands falls within a relatively narrow band of positivity and will not have much impact in deducing the popularity of a particular post posted by brands.

Therefore, based on this conclusion this feature was not merged with our contextual and visual dataset as it was concluded that sentiment score for all the three brands lies within the same range of positivity and will have minimal impact on training our model to improve the accuracy and resultant of model in predicting the popularity of brands.

Conclusion

This research presents a new multi-label multimodal, trained with multioutput for popularity prediction of top 3 beauty brands – Lakme, L'Oreal and Maybelline. The proposed method includes statistical texture feature extraction from images for each brand using skimage module features such as Local Binary Pattern (LBP), Gray Level Co-occurrence Matrix (GLCM) and Gabor Filter. The method also includes contextual and social feature along with image features such as likes count, comment count for each image, dimensions of an image, following and labels which altogether forms the complete dataset for our analysis to be carried out. The proposed method comes to the conclusion that, when compared to other baseline machine learning models like Random Forest Regressor, Support Vector Regressor (SVR), and Linear Regression, the performance of the XGBoost model—both the DMatrix and Regressor—performs best on the provided dataset. The performance of popularity prediction was measured using the metric known as RMSE, and XGBoost came out on top with a score of 0.03952. Furthermore, additional sentiment analysis study was carried out on textual data taken from captions, and it was determined that each brand's sentiment polarity falls within a similar range and will have no impact on the accuracy or performance of popularity prediction, if it is included. The final outcome of the predicted numbers of likes associated with each brand's (label) demonstrated that Maybelline's popularity greatly outpaced that of L'Oreal and Lakme.

Limitation and Future Work

1. Several studies have been conducted in this multimodal popularity prediction where they have utilized deep learning and neural network techniques for feature extraction. However, deep learning features though quite popular but it requires lots of data and high-performance GPUs. As there was limit to data extraction with our newly introduced dataset of three beauty brands, this led to a constraint with which we were unable to proceed with the deep learning models. Therefore, we learned about the best combination of feature that could be extracted for better prediction.
   In future, we aim to find the means to scrap a larger dataset with which can perform deep learning model techniques.

2. In our research, we have taken into account a limited set of features which are crucial for popularity prediction. However, there are few external and internal features such time and date of post, comments posted by users, number of verified accounts etc, which could be beneficial in performance enhancement of our model and help in predictability.
   In future, we aim to work on these features to illustrate their impact on Multimodal popularity prediction.

3. In Addition, we also aim to analyse in the future that how location and culture-based factors can affect the popularity of images which was not possible given time constraint.

4. From a data science point of view, the limits to predictability of human behaviour is a challenging research question. Hence, understanding the human behaviour with deep sentimental analysis on comments and then identifying hate or appraising speech, that be included as a feature to grab a better understanding in the popularity prediction.

References:

1) Pinto, H., Almeida, J.M. and Gonçalves, M.A., 2013, February. Using early view patterns to predict the popularity of youtube videos. In Proceedings of the sixth ACM international conference on Web search and data mining (pp. 365-374).

2) Khosla, A., Das Sarma, A. and Hamid, R., 2014, April. What makes an image popular?. In Proceedings of the 23rd international conference on World wide web (pp. 867-876).

3) Almgren, K., Lee, J. and Kim, M., 2016, October. Prediction of image popularity over time on social media networks. In 2016 Annual Connecticut Conference on Industrial Electronics, Technology & Automation (CT-IETA) (pp. 1-6). IEEE.

4) Gelli, F., Uricchio, T., Bertini, M., Del Bimbo, A. and Chang, S.F., 2015, October. Image popularity prediction in social media using sentiment and context features. In Proceedings of the 23rd ACM international conference on Multimedia (pp. 907910).

5) Aloufi, S., Zhu, S. and El Saddik, A., 2017. On the prediction of flickr image popularity by analyzing heterogeneous social sensory data. Sensors, 17(3), p.631.

6) Almgren, K., Lee, J. and Kim, M., 2016, August. Predicting the future popularity of images on social networks. In Proceedings of the The 3rd Multidisciplinary International Social Networks Conference on SocialInformatics 2016, Data Science 2016 (pp. 1-6).

7) Paolanti, M., Kaiser, C., Schallner, R., Frontoni, E. and Zingaretti, P., 2017, September. Visual and textual sentiment analysis of brand-related social media pictures using deep convolutional neural networks. In International conference on image analysis and processing (pp. 402-413). Springer, Cham

8) Garg, M. and Dhiman, G., 2021. A novel content-based image retrieval approach for classification using GLCM features and texture fused LBP variants. Neural Computing and Applications, 33(4), pp.1311-1328.

9) Allagwail, S., Gedik, O.S. and Rahebi, J., 2019. Face recognition with symmetrical face training samples based on local binary patterns and the Gabor filter. Symmetry, 11(2), p.157.

10) Alharan, A.F., Fatlawi, H.K. and Ali, N.S., 2019. A cluster-based feature selection method for image texture classification. Indonesian Journal of Electrical Engineering and Computer Science, 14(3), pp.1433-1442.

11) Barburiceanu, S., Terebes, R. and Meza, S., 2021. 3D texture feature extraction and classification using GLCM and LBP-based descriptors. Applied Sciences, 11(5), p.2332.

12) Ding, K., Wang, R. and Wang, S., 2019, October. Social media popularity prediction: A multiple feature fusion approach with deep neural networks. In Proceedings of the 27th ACM International Conference on Multimedia (pp. 2682-2686).

13) Abousaleh, F.S., Cheng, W.H., Yu, N.H. and Tsao, Y., 2020. Multimodal deep learning framework for image popularity prediction on social media. IEEE Transactions on Cognitive and Developmental Systems, 13(3), pp.679-692.

14) Gupta, V., Jung, K. and Yoo, S.C., 2020. Exploring the power of multimodal features for predicting the popularity of social media image in a tourist destination. Multimodal Technologies and Interaction, 4(3), p.64.

15) Hu, J., Yamasaki, T. and Aizawa, K., 2016, May. Multimodal learning for image popularity prediction on social media. In 2016 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW) (pp. 1-2). IEEE.

16) Chen, J., Liang, D., Zhu, Z., Zhou, X., Ye, Z. and Mo, X., 2019, October. Social media popularity prediction based on visual-textual features with xgboost. In Proceedings of the 27th ACM International Conference on Multimedia (pp. 2692-2696).

17) Zhu, C., Cheng, G. and Wang, K., 2017. Big data analytics for program popularity prediction in broadcast TV industries. Ieee Access, 5, pp.24593-24601.

18) Folgado, M.G. and Sanz, V., 2022. Exploring the political pulse of a country using data science tools. Journal of Computational Social Science, 5(1), pp.987-1000.

19) Hasan, M.R., Maliha, M. and Arifuzzaman, M., 2019, July. Sentiment analysis with NLP on Twitter data. In 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2) (pp. 1-4). IEEE.

20) Adikari, A. and Alahakoon, D., 2020. Understanding citizens’ emotional pulse in a smart city using artificial intelligence. IEEE Transactions on Industrial Informatics, 17(4), pp.2743-2751.

21) Gharge, S. and Chavan, M., 2017, March. An integrated approach for malicious tweets detection using NLP. In 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 435-438). IEEE.

22) Garg, R., Kiwelekar, A.W., Netak, L.D. and Ghodake, A., 2021. i-Pulse: A NLP based novel approach for employee engagement in logistics organization. International Journal of Information Management Data Insights, 1(1), p.100011.

23) Humeau-Heurtier, A., 2019. Texture feature extraction methods: A survey. IEEE access, 7, pp.8975-9000.

24) Meghawat, M., Yadav, S., Mahata, D., Yin, Y., Shah, R.R. and Zimmermann, R., 2018, April. A multimodal approach to predict social media popularity. In 2018 IEEE conference on multimedia information processing and retrieval (MIPR) (pp. 190-195). IEEE.

25) Nguyen, L.T.K., Chung, H.H., Tuliao, K.V. and Lin, T.M., 2020. Using XGBoost and skip-gram model to predict online review popularity. SAGE Open, 10(4), p.2158244020983316.

26) Li, L., Situ, R., Gao, J., Yang, Z. and Liu, W., 2017, October. A hybrid model combining convolutional neural network with xgboost for predicting social media popularity. In Proceedings of the 25th ACM international conference on Multimedia (pp. 1912-1917).

27) Nisa, M.U., Mahmood, D., Ahmed, G., Khan, S., Mohammed, M.A. and Damaševičius, R., 2021. Optimizing Prediction of YouTube Video Popularity Using XGBoost. Electronics, 10(23), p.2962.

28) Sanjo, S. and Katsurai, M., 2017, November. Recipe popularity prediction with deep visual-semantic fusion. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 2279-2282).

29) Pouyanfar, S., Wang, T. and Chen, S.C., 2019, March. A multi-label multimodal deep learning framework for imbalanced data classification. In 2019 IEEE conference on multimedia information processing and retrieval (MIPR) (pp. 199-204). IEEE.

30) Kim, W., 2020. Comparative Study of Tokenizer Based on Learning for Sentiment Analysis. Journal of the Korean Society for Quality Management, 48(3), pp.421-431.
31)Shahana, P.H. and Omman, B., 2015. Evaluation of features on sentimental analysis. Procedia Computer Science, 46, pp.1585-1592.

32) Daugman, J.G.: Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. Journal of the Optical Society of America 2(7), 1160–69 (1985)

33) Vazquez-Fernandez, E., Dacal-Nieto, A., Martin, F. and Torres-Guijarro, S., 2010, June. Entropy of gabor filtering for image quality assessment. In International Conference Image Analysis and Recognition (pp. 52-61). Springer, Berlin, Heidelberg.

34) Gujjar, J.P. and Kumar, H.P., 2021. Sentiment analysis: Textblob for decision making. Int. J. Sci. Res. Eng. Trends, 7(2), pp.1097-1099.

35) Laksono, R.A., Sungkono, K.R., Sarno, R. and Wahyuni, C.S., 2019, July. Sentiment analysis of restaurant customer reviews on TripAdvisor using Naïve Bayes. In 2019 12th International Conference on Information & Communication Technology and System (ICTS) (pp. 49-54). IEEE.

36) Elbagir, S. and Yang, J., 2019, March. Twitter sentiment analysis using natural language toolkit and VADER sentiment. In Proceedings of the international multiconference of engineers and computer scientists (Vol. 122, p. 16).

37) Hutto, C. and Gilbert, E., 2014, May. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 216-225).

38) Dangeti, P., 2017. Statistics for machine learning. Packt Publishing Ltd.

39) Schapire, R.E. and Freund, Y., 2012. Foundations of machine learning.

40) https://medium.com/sfu-cspmp/xgboost-a-deep-dive-into-boosting-f06c9c41349

41) https://xgboost.readthedocs.io/en/stable/parameter.html

42) Varela-Rodríguez, M. and Vicente-Mariño, M., 2020, October. Automated image extraction from Instagram for social research: A technical and ethical exploration. In Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality (pp. 588-592).

43) Akrianto, M.I., Hartanto, A.D. and Priadana, A., 2019, November. The Best Parameters to Select Instagram Account for Endorsement using Web Scraping. In 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE) (pp. 40-45). IEEE.

44) Sharma, K., Narang, Y., Gupta, C., Teotia, Y. and Bhardwaj, P., 2020. Clustering emotions of people in various Indian cities using Instagram. Journal of Statistics and Management Systems, 23(7), pp.1199-1211.

45) https://github.com/instaloader/instaloader ((The MIT License) Copyright (c) Alexander Graf and André Koch-Kramer.

46) https://towardsdatascience.com/predicting-the-popularity-of-instagram-postsdeeb7dc27a8f

47) S. Rogers and M. Girolami, A first course in machine learning, CRC Press, 2016.

48) K. P. Bennett and C. Campbell, "Support vector machines: hype or hallelujah?," Acm Sigkdd Explorations Newsletter, vol. 2, no. 2, pp. 1-13, 2000

49) T. K. Ho, "Random Decision Forests," in Proceedings of the Third International Conference on Document Analysis and Recognition, 1995. [68] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, p. 5–32, 2001

50) https://medium.com/@theclickreader/random-forest-regression-explained-withimplementation-in-python-3dad88caf165

51) https://medium.com/coinmonks/support-vector-regression-or-svr-8eb3acf6d0ff

APPENDIX A

Git Repository provide by University of Birmingham:
https://git-teaching.cs.bham.ac.uk/mod-msc-proj-2021/sxx182

The above-mentioned repository consists of the following.

1) Dataset for the all the three brands – Lakme, Loreal and Maybelline including images, json and Text.
2) Two Jupyter notebooks - The .ipynb files, one referring to popularity prediction using image and json and the other for sentiment analysis using textual data.
   - Multimodal_popularity_prediction.ipynb
   - Textual_Data_sentiment.ipynb

Steps to executing the code:
Multimodal_popularity_prediction.ipynb

1) Skip the below steps while running the program
   - cell [3][4] of the program which split and label the dataset
   - cell [6] of the program as it splits the dataset into train and test.
   - cell [46] as it again split the dataset of json files into train and test.

2) Change the folder path defined in cell 5, 7, 45 & 47 as per your own system directory/path where the dataset will be downloaded.

   The above steps are the key to success of the Jupyter notebook "Multimodal_popularity_prediction.ipynb"

Textual_Data_sentiment.ipynb

1) Change the folder path defined in cell 13 as per our own system/directory path where the dataset will be downloaded.

Note: The model's hyperparameter has been tuned using GridSearchCV, due to which it may take longer time for execution. Also, image feature extraction takes around an hour for training and testing dataset which is normal and is not alarming.