Emerging Trends In Business Analytics And Decision Sciences

Effect Of Image Usage in Social Media: A Study of the Relationship between Higher Attributes Of Images In Social Media Posts And Likes, Comments And Shares Garnered By Them

Medhavi Bhardwaj [1], Prof. (Dr.) Arpan Kar [2]

[1] Indira Gandhi Delhi Technical University for Women, James Church, New Church Rd, Opp. St, Kashmere Gate, New Delhi, Delhi 110006

[2] Indian Institute of Technology Delhi, IIT Delhi Main Rd, IIT Campus, Hauz Khas, New Delhi, Delhi 110016

Abstract:

Purpose: The purpose of this paper is to examine the relationship between the social media engagement garnered by a post, i.e. number of likes, comments and shares it receives and the higher attributes of images utilized in the post. The purpose of this paper is to study and closely examine the relationship between higher attributes of images in Facebook posts and social media engagement garnered by them in the form of likes, comments and shares. These statistics can help the companies analyze the trends in customer preferences and customize products to increase revenue.

Methodology: The model has been developed through multimodal analysis, by extracting higher attribute values from a dataset of 300 Facebook posts from Fortune 500 companies, with images using deep learning models, along with their likes, comments and shares. The relation was derived using models like linear regression and negative binomial regression.

Findings: The study indicates that the higher attributes of images (e.g. semantic segmentation, object recognition, similarity, etc.) influence the amount of engagement a social media post with an image gets, to some extent. Further machine learning algorithms were used to test the validity and robustness of the relation.

Originality: The research is one of the few to further study how the selection of images for posts based on their higher attributes can help in increasing customer engagement. These data statistics and relations can help companies and industries identify the preferences of the target audience better, which can help them generate customized contents and create tailored products resulting in higher revenue and customer satisfaction. The research can be applied to areas of marketing, advertisement, deployment and testing and sales managements of new products developed by a company.

Keywords: Social media engagement, deep learning, machine learning, image processing

1 Introduction

The world has undergone drastic changes in the last century. Progress in technology has altered the methods of communication, making them vastly different from the ones used earlier. In the

traditional economy, a market was confined to a physical space, limited in terms of space, time and accessibility. The introduction of the Internet has increased the accessibility of the market (through e-commerce and online shopping) to the extent that it can be accessed any time and from anywhere as long as there is an electronic device and a stable internet connection. Social media which originally began as a platform for human interaction, now has been developed into a multipurpose platform that can be used for brand positioning, advertising and many others aspects of marketing management. For any business, it is imperative to keep up with the advancements taking place in the world of marketing and tools used in it (Ravı & Kumar, 2021). Social media platforms have become a prominent part of online communication. Companies compete for consumers' attention in order to increase sales of their products. Customers also participate in meaningful interactions with brands and communities formed by them, beyond just one purchase (Brodie, Ilic, Juric, & Hollebeek, 2013). Thus, the relation created between customers and brands is formed at a more intimate level (Sashi, 2012). Social media platforms such as Facebook, WhatsApp, Twitter, etc. facilitate free non-restrictive expression and allow informal and formal methods of communication. Thus, they are popular communication platforms used by customers worldwide. Thus, these platforms, with large viewings and a large number of users are an ideal place for businesses and companies to advertise their products and find a large target audience. Exponential growth has occurred in the use of social media platforms such as WhatsApp, Instagram, and Facebook over the past decade (Chen and Qasim, 2021).

Brands must keep track of trends and observe people's interests closely in order to understand the customer profile. This helps them maintain a competitive edge over their peers in the industry.
The study conducted by Jamil and Dunnan, et. al. confirmed that most administrators are concerned with the influence of brand community management in creating business advantage (Jamil, Dunnan, et. al., 2022). According to this study, if a company can successfully assist users to easily identify with a particular brand community, strong relationships will be fostered between the consumers and the brand, hence creating customer's loyalty (Ebrahim, 2020). Chi (2011) defines social media marketing as a "connection between brands and consumers, while offering a personal channel and currency for user centered networking and social interaction." Social media marketing methods have also changed over time. Hence companies have to use social media in a way which is consistent with their business plans and can help them achieve their targets in terms of sales and profits (Mangold and Faulds 2099).
The advertisement of products is done using a variety of media in posts including images, audio clips, videos, embedded links to websites and other videos, in social media posts.

These days, consumers gain a new role with social media, becoming 'content creators' and thus, functional consumers instead of just consuming, as in the past. There are many social media applications or tools that facilitate this such as blogs, micro blogging applications (such as Twitter), social networking sites (such as Facebook), podcasts, and video and photo sharing sites (such as YouTube and Flickr). Given this reality, companies, especially marketers, may find it highly useful to integrate social media into marketing and their marketing strategies (Nadaraja, Rubathee, et. al., 2013)

Usually, people process images more easily than text, as the brain is a visual learner. It stores most of the information processed by the body in the form of memories which are similar to film,

which is a series of images viewed at rapid speeds. The brain associates terms and clusters of information with particular memories, and thus remembers them. In today's world, where people have remarkably short attention spans. According to a research study conducted by Microsoft in 2015, the average span of attention for a human is 8 seconds which is even shorter than that of a goldfish, which is 9 seconds (National Center for Biotechnology Information, 2015). Using the relevant images helps in increasing engagement, gain attention, build up loyalty, and leave a long-lasting impression. Images can be used to tell stories in a quick glimpse and make things easier to remember in comparison to long paragraphs with endless details. According to the Digital Marketing Agency In Oxford, images in particular are used by companies to increase shareability, and increase engagement.

In this paper, we examine some of the higher attributes of images used in social media posts, and how they are related to the amount of social media engagement each post generates.

This paper aims to answer the following questions:

1.) Do higher attributes of images used in social media posts influence the amount of engagement (number of likes, shares or comments a post gets on a social media platform (e.g. Twitter, Facebook, Instagram, etc.)?
2.) How are higher attributes of images related the amount of engagement a post gets?
3.) Which higher attributes are related to the level of engagement a social media post gets, and how do they influence this engagement

2. Literature Review

Research done in social media engagement previously examines a variety of attributes of images including resolution, image quality, color scheme, etc. The main characteristics influencing a photo's performance in social media include, among others, visual appeal, relevance, quality, emotional resonance, brand consistency, composition, storytelling.

Weick (1995) investigates the concept of theorizing, taking reference from Sutton and Staw's discussion of five different article parts. Merton (1967) suggests that approximation can take any one of the four forms: 1)general orientations in which broad frameworks specify types of variables that should be taken into account, without explaining the relationships among them; 2) Analysis of concepts, including their clarifications, definitions but they may not be interrelated; 3) post factum interpretation where ad hoc hypotheses are derived from a single observation, and no efforts are made to explore alternate explanations, or new observations; 4) empirical generalization where an isolated proposition summarizes the relationship between 2 variables, but further interrelations are not attempted. Runkel and Runkel's (1984) state that theory is a continuum rather than a dichotomy.

Theory belongs to the same family of words that includes guess, speculation, proposition, and conjecture. The word theory can be used for a wide range of things from 'a simple guess' to 'a system of assumptions and accepted principles and rules of procedure devised to analyze, predict, or otherwise explain the nature or behavior of a specified set of phenomena', as defined in the American Dictionary. According to Karl E. Weick, the term theory can be used during various stages of the process of theorizing, rather than just using it to label the final conclusion, which is arrived at.

It is often very difficult to sort out the actual theory from amateur attempts while working with people with varying levels of knowledge and expertise. Further the need to properly paraphrase and explain the reason for each citation or reference made either in books, research papers or any other scholarly articles is also emphasized, so that readers can easily understand why the

particular reference has been utilized. So, according to Karl, if the references are somehow connected then we approach something closer to actual theory. Data itself isn't actual theory according to Bacharach (Academy of Management Review,1989). On the other hand, Starbuck (1983) also argues that theorists can make various prescriptions based on data alone without bringing theory in the middle, similar to the way the best doctors treat symptoms directly without relying on the diagnosis for determining the treatment. In both cases, theories and diagnosis provide a summary of relations observed between treatments/prescriptions and symptoms or data. Also, since combinations of data and observable characteristics are more, there will be a larger number of conclusions and relationships between them. Using theories to condense information may lead to loss of data or essential factors which must be taken into account while drawing conclusions. This leads to random errors being injected into the conclusions drawn. Starbuck (1993) has given a summary of the argument in this way: academic research follows a model similar to that of medical schools, where the scientists try to translate data, which are like symptoms into theories. So, theories are diagnoses and prescriptions are like treatments in this case.

So, organizations have been compared to complex human bodies and theories may not necessarily capture all the information or the magnitude or scale to which a derived conclusion can be applied. Neither can theories determine unique prescriptions. Hence, they may not necessarily be able to provide accurate solutions while taking into account various factors of a problem or a situation.

Previously, many studies have conceptualized social media engagement in terms of linguistic content. However, these days, businesses have utilized new ways of connecting and communicating with their target audience- the customers, by using a combination of linguistic and visual content. This wide variety of content has made researchers start analyzing it through multimodal analysis (Shao and Janssens, 2022).

Juan Caballero, Gibran Gomez, et al. suggested GoodFATR, a automated digital platform which collects reports of threats from multiple sources, finding and extracting indicators of compromise(IOCs) from them. There are 6 main sources from which the data is collected: RSS, Twitter, Telegram, APTNotes, Chain Smith and Malpedia. They are constantly monitored, and threats are downloaded

Singh, Gandhi and Kar et al. (2023) studied effects of social media image content in big business firms and companies to increase the social media engagement, which emphasises the importance of strategically designing content in form of images as a marketing strategy. Creating social media posts that engage the audience is the best way to optimize social media use to maximise outreach to more customers. However, according to Benton, about 60% of business to business (B2B) content creators acknowledge it to be one of the biggest challenges (Benton,2017). A computation extensive research model was designed based on the stimulus organism response (SOR) theory. The study used 39129 Facebook posts from 125 companies selected out of the Fortune 500 firms list. Attributes from both images and text were measured using deep learning models. An inferential analysis is used based on the least squares regression. The other machine learning algorithms were used to analyse strength and sound architecture of the proposed model and determine whether it was a valid and sound source for prediction of engagement metrics. A few prominent examples include: the k-nearest neighbour, support vector regression, random forest, decision trees.

The findings of the undertaken study indicated that the social media image content posted by big business firms had a significant impact on their social media engagement. The visual and

linguistic attributes of images were extracted using deep learning methods and models and the distinctive effect of each feature was verified empirically. This study offers many practical insights which have been formed by observing various online marketing methods such as embedded marketing, advertising with the help of image processing and statistical information of social media.

Social media (websites, platforms, applications, etc.) allow companies to increase the spread of information by sharing their knowledge and information with users and customers (Wukich,2022).

It helps the firms by offering an insight to understand the rapidly evolving needs of the customer market so that they can provide rapid responses accordingly. The main aim of organizations is to reach the full potential of social media engagement to increase their sales, achieve full customer satisfaction, and increase the quality of decision making within the company (Rutter, Barnes, Roper, Nadeau, and Lettice, 2021; Simon and Tossan, 2018). Gandhi and Kar's study aims to answer 2 main questions:

1) Why is social media engagement influenced by attributes of images in social media posts?

2)How do an image's linguistic and visual attributes in particular affect its engagement (the number of likes, comments and shares a post with an image receives) on social media?

They have provided a conceptual model which evaluates firm generated content using the SRM lens to verify the effects on social media engagement, using multimodal analysis. Computer vision ad natural language processing algorithms were used to extract and evaluate visual, linguistic and a combination of both attributes from the images in the social media content posted by the firms. The findings suggest that visual features of an image (including presence of human faces, quality, entropy of an image) and the textual content in the image are big drivers and factors influencing social media engagement. Machine learning algorithms are used to perform multimodal analysis to do a comparative analysis (Han, Lam, Zhan, Wang, Dwivedi, and Tan, 2021). Inferential analysis was the main method used. The results suggested that presence of only a human face in a picture had a significant positive effect on the engagement.

Semantic segmentation can be measured in terms of highest percentage of Overall Pixel (OP) accuracy, which measures the proportion of correctly labelled pixels. The Per-Class (PC) accuracy measures the proportion of correctly labelled pixels for each class and then averages over the classes. Therefore, the background region absorbs all false alarms without affecting the object class accuracies. This measure is suitable for datasets with no background class. These scores are common to the MSRC dataset (Shotton, Winn, et. al., 2006).

Image similarity measures play an important role in image fusion algorithms and applications, such as duplicate product detection, image clustering, visual search, change detection, quality evaluation, and recommendation tasks. These measures essentially quantify the degree of visual and semantic similarity of a pair of images. In other words, they compare two images and return a value that tells you how visually similar they are. Ono and Mattmann, et. al. (2022) developed a similarity search ability which returns a list of images that are semantically similar to a query image provided by users. It works by extracting an intermediate layer from the VGG19 image encoding part of SCOTI (Conv5_3). It then compares the feature vector against the feature vector of other images using the cosine distance metric to find the most similar or dissimilar (in the case of newly searched) images. The opencv method in python uses histogram based approaches. Histograms capture the distribution of pixel values in an image. By comparing the histograms of two images, you can measure their similarity.

The Histogram Intersection and Histogram Correlation metrics are commonly used for this purpose. Python's opencv library also provides requisite tools to create, compute and compare histograms.

Using pretrained deep learning models is another way to extract features and thus find out similarity score between a pair of images. The similarity between images can then be computed based on the cosine similarity or Euclidean distance of these feature vectors. To improve the accuracy, we can preprocess the images.

MIT has created neural networks, based on our understanding of how the brain works, that allow software to identify objects almost as quickly as primates do. Average Precision (AP) and mean Average Precision (mAP) are the most popular metrics used to evaluate object detection models, such as Faster R_CNN, Mask R-CNN, and YOLO, among others. The same metrics have also been used to evaluate submissions in competitions like COCO and PASCAL VOC challenges.

Machine learning models use various features which are crucial components which enable them to make accurate predictions, by understanding data patterns and processing data accurately (Keco, Obucic, et. al., 2024). The digital marketing strategy has become increasingly more important and popular as companies aim to increase social media users' engagement, revenue and brand awareness. The objective of the study conducted by Dimitris C. Gkikas and Prokopis K. Theodoridis (2024) was to calculate the text characteristics of organic photo posts like: text readability, hashtags number and characters number. Using data mining classification models, it examines whether these characteristics affect organic post user engagement for lifetime post engaged users and people who have liked a page and engaged with a post in a lifetime. The findings of the study revealed how post texts' content characteristics impact performance metrics thus helping the marketers to better formulate their social media organic strategies, the company to increase impressions, reach and revenues, and the customers to comprehend the post message and engage with the brand (Dimitris C. Gkikas & Prokopis K. Theodoridis, 2024) . Machine learning algorithms are categorised into two basically supervised and unsupervised techniques:

a) Supervised learning: Applications in which the training data comprise example of the input vectors along with their corresponding target vectors are known as supervised learning methods (Lakshmi, 2016).
b) Unsupervised learning: In other pattern-recognition problems, the training data consist of a set of input vectors x without any corresponding target values. The goal in such unsupervised learning problems may be to discover groups of similar examples within the data (Manar and Stephane, 2015).

Supervised techniques in machine learning areas can be further grouped as below:

1.Classification: This assigns a category to each object such as OCR, text classification, speech recognition.

2.Regression: This is used to predict a real value for each object such as stock prices, values, economic variables and ratings.

3.Clustering: This is based on partition data into homogeneous groups (analysis of very large datasets).

4.Ranking: The order objects according to some criterion (relevant web pages returned by a search engine).

5.Dimensional reduction: To find lower-dimensional manifold preserving ties of the data (computer vision).

6.Density estimation: This is used for learning probability distribution according to which

data have been sampled.

3.Conceptual Model

3.1Models used in social media analysis

Trunfio and Rossi(2021) conducted a study, where out of all the prior literature on social media engagement, only a small part (10% of the total papers analysed in the study) explore the topic of social media engagement through a perspective which is purely theoretical or conceptual. Many customer engagement conceptualisations have been proposed in the literature, based on various theoretical backgrounds, in particular service-dominant logic, and relationship marketing. From a psychological perspective, one of the first definitions of customer engagement was given by Bowden (2009) that visualises it as a psychological process driving customer loyalty. Similarly, Brodie et al. (2011) have defined customer engagement as a psychological state that occurs due to interactive, co-creative customer experiences devised with a focal point or objective. In later works, while focusing on the behavioural aspects, it has been defined as the intensity or level of an individual's participation in an organisation's offerings or organisational activities (Vivek et al., 2012). More recently, from a value-based perspective, customer engagement has been defined as the mechanics that customers use to add value to the firm (Kumar et al., 2019).

3.1.1The Stimulus and Response Model

Gandhi, Kar, et al. (2023) made use of the stimulus response model to show how multimodality can be used to analyse the relationship between attributes of social media posts and engagement generated by them. This model is an extension of a predetermined pattern of behaviour, which has been used to make an attempt to understand and explain certain reactions to stimuli.

3.2Research Conceptual Model Using Hypothesis development

Various perspectives from theory and a closer look at previous work done in the area of social media analysis reveal facts which helped us design the following conceptual model.

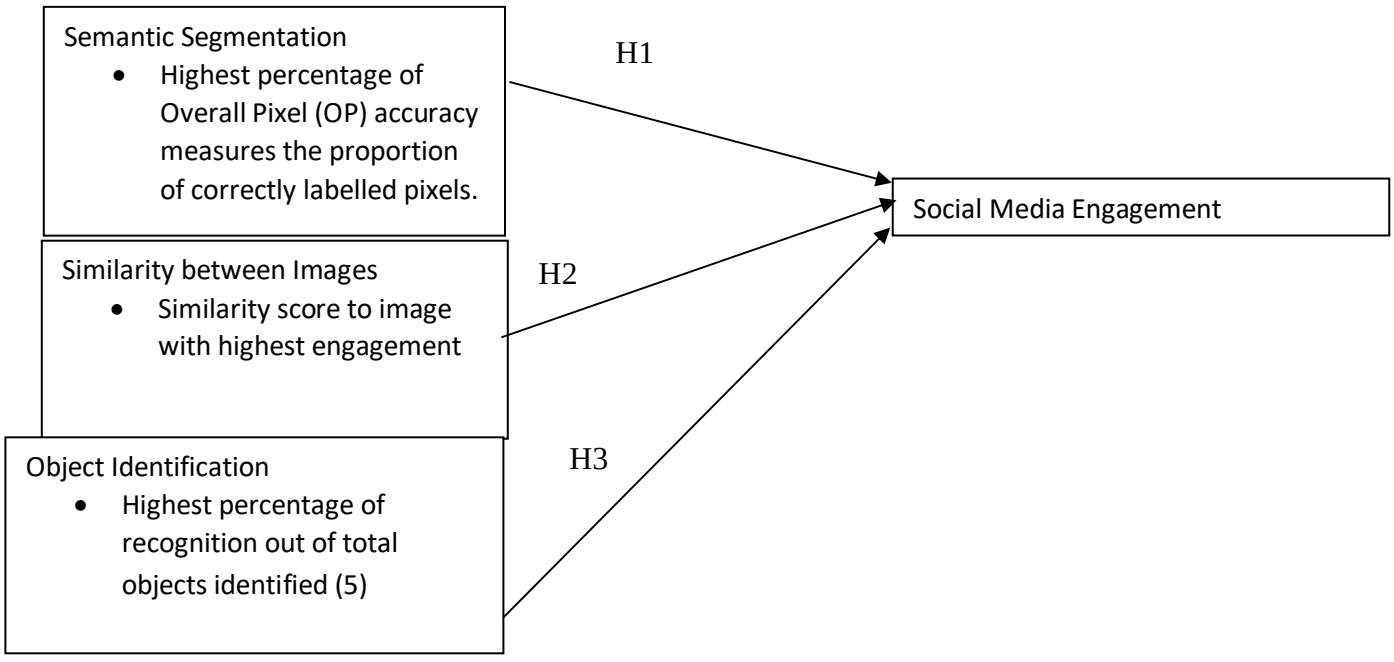

Fig 1 : Conceptual model

2.1 Semantic Segmentation

According to the study conducted by Rathnayake and Ntalla (2020) there are many approaches to analyzing images that exceed the boundary of meaning-making. For instance, notions such as visual clutter—"state in which excess items, or their representation or organization, lead to a degradation of performance at some task" (Rosenholtz et al., 2007). Rathnayake and Ntalla (2020) propose visual affluence as a related measure that can be applied across different types of visual content and sample sizes, from a few images to large volumes of images. They discuss how image segmentation can be used to develop a visually oriented basis for classifying images. Further work on implications of visual affluence can benefit a range of disciplines beyond social media studies, such as advertising, branding, and political communication.

Hypothesis 1a) A higher level of semantic segmentation leads to a higher level of social media engagement.

Social media engagement in this context means likes, comments and shares on posts containing images.

3.2.1 Similarity between Images

Li and Xie(2020) found that professionally taken pictures tend to boost the number of retweets in comparison to pictures taken by amateur photographers. Existent research (Hagtvedt and Patrick 2008; Zhang et al. 2017) shows that high-quality images can improve user engagement on social media posts, as shown by their significant and positive effect on the number of retweets.

Hypothesis 1b) Images incorporating similar styles or those that are similar to each other will have the same level of social media engagement

For instance, social media posts with images of famous personalities and celebrities are likely to get a higher number of likes, comments and shares. Hence, product companies, in particular fashion companies select celebrities with a very large following as their rand ambassadors to promote their products.

3.2.2Object Identification

Zhang, Lee, et. al. (2017) conducted a study to predict social media engagement by using computer vision. Using Google Vision AI, they tried to identify objects in images in social media posts about food, and determine the relation between the objects reflected in the image and the likes or engagement received by the post. This research found that the confidence score given to food objects (similar to food typicality as confirmed by human coders) is positively related to engagement received by the image post. This finding was replicated in an experiment, which showed evidence that this effect can be explained through the following reason: typical-appearing foods are easier to mentally process and thus elevate positive affect in comparison to more unique, atypical-appearing foods.

Hypothesis 1c) Higher percentage of more identifiable objects will have a positive impact on the social media engagement.

The parameter of number of objects identified can be utilized.

Several customer engagement conceptualisations have been proposed in the literature, drawing on various theoretical backgrounds, particularly service-dominant logic, and relationship marketing. From a psychological perspective, one of the first definitions of customer engagement is the one of Bowden (2009) that conceptualises it as a psychological process that drives

customer loyalty. Similarly, Brodie et al. (2011) define customer engagement as a psychological state that occurs by interactive, co-creative customer experiences with a focal object. Later, focusing on the behavioural aspects, it has been described as the intensity of an individual's participation in an organisation's offerings or organisational activities (Vivek et al., 2012). Tan and Lim(2020) have examined the importance of social media engagement rate as a metric to measure social media engagement of various firms.

4. Research and Methodology

In previous research studies undertaken throughout the decades, usually data is collected, cleaned, analyzed and then processed, in order to get some output. On the basis of this output, new insights are gleaned and applied to practical fields of application. Features can be defined as follows: "Each individual, independent instance that provides the input to machine learning is characterised by its values on a fixed, predefined set of features or attributes" (Witten & Frank, 2002). In machine learning, particular features or attributes of the data are selected in order to process the input data on their basis and give the requisite output. This becomes important, especially when determining the relations between various entities, and finding out how their attributes are interrelated to each other. It also helps in finding out correlations.

Furthermore, one may also derive more complicated features by using raw data. A machine learning model may extract features that represent the interaction between multiple features or features that capture the overall structure or patterns of the data. Another possibility is that the model may extract features that represent the overall structure of the data.

Table 1. Feature type and description

Numeric features	Numeric features represent numeric values, such as the number of items in a set
Binary features	Binary features have only two possible values, such as 0 and 1.
Categorical features	Categorical features represent categories or classes, such as different types of cats
Continuous features	Continuous features continuously change values, like humidity measure.
Ordinal features	Ordinal features have a defined order, such as scale from 1 to 5

Source: (James et al., 2013); (Murphy, 2018); (Bishop, 2006).

4.1 Numeric (Quantitative) Features

Numeric features measure various quantities numerically. Statistical analysis and most computer software operations require numeric data as input. These numeric values are a method for analyzing the various properties of data being studied.

Parameters of accuracy

The mean squared error metric indicates the average squared difference between the observed actual outcomes and the outcomes predicted by the model. A lower MSE indicates better fit. A higher value. The R^2 value indicates how well the model explains the variability of the response data around its mean. In this case, the negative R^2 suggests that the model does not fit the data well and performs worse than a horizontal line representing the mean of the dependent variable. indicates a worse fit for the data fed to the machine learning model.

4.2 Methodology

Conducting a study based on data exploration and deriving new insights from it, typically includes: data collection and cleaning, feature selection and extraction and using machine learning models to generate predictions, and then evaluating the predictions on the basis of the parameters like accuracy, by comparing them with the original predictions, and finally deriving insights form these predictions.

4.2.1 Data Collection and Cleaning

For this study, a dataset of most recent Facebook posts was extracted from the official accounts of Fortune 500 companies like Amazon, Walmart and Berkshire Hathways using Apify. The dataset was exported to Excel, with attributes of the number of likes, comments and shares. Using Python libraries for data handling like seaborn, scikit learn, NumPy, and Pandas, data was stored in the form of a Pandas dataframe and cleaned. Massive amounts of data are available for the organization which will influence their business decision. Data cleansing process mainly consists of identifying the errors, detecting the errors and corrects them. Vast amounts of data need to be analyzed quickly, the data cleansing process is complex and time-consuming in order to make sure the cleansed data have a better quality of data(Ridzuan, Fakhitah, et. al.,2019). When processing some of the even larger datasets, handling a sophisticated mechanism to discover errors or managing large arbitrary errors, the overhead of data cleansing may reach up to more than 60% of the data scientists' time (Crowdflower, 2016). Wang, et al. (2014) have designed and proposed Cleanix; a parallel big data cleansing system aims to solve the issue related to the volume and variety of big data. Four types of data quality issues are tackled by Cleanix which are abnormal value detection, incomplete data filling, deduplication, and conflict resolution. It is developed with the scalability, unification and usability features which enable Cleanix to perform data cleansing and data quality reporting task in parallel. Using bad or poor data in a BI or data analysis process can lead to incorrect analysis, business operation errors, and bad business strategies. Addressing bad data before it’s executed in a data analysis process saves businesses money by reducing the expense of fixing bad data results after the data is processed, including the added cost of interrupting business operations to correct the results of bad data. The data cleaning steps are:

- Remove irrelevant data
- Deduplicate redundant data
- Repair structural errors
- Address missing data
- Filter out data outliers
- Validate that the data is correct

Database normalization is a database design principle that helps you create database tables that are structurally organized to avoid redundancy and maintain the integrity of the database.

4.2.2Measurement and extraction of Higher attributes of Images

Automatic attribute discovery methods have gained a lot of popularity to extract sets of visual attributes from images or videos for various tasks. Despite their good performance in some classification tasks, it is difficult to evaluate whether the attributes discovered by these methods are meaningful and which methods are the most appropriate to discover attributes for visual descriptions. In its simplest form, such an evaluation can be performed by manually verifying whether there is any consistent identifiable visual concept distinguishing between positive and negative exemplars labelled by an attribute. This manual checking is tedious, expensive and labour intensive. Liu, Wiliem, et. al. (2017) proposed a novel attribute meaningfulness metric to tackle this problem, with which automatic quantitative evaluation can be performed on the attribute sets; thus, reducing the enormous effort to perform manual evaluation. The proposed metric is applied to some recent automatic attribute discovery and hashing methods on four attribute-labelled datasets. A user case study was further conducted to validate the result. Using the Python libraries for machine learning-OpenCV, Pytorch and Tensorflow higher attributes of images were extracted, having numerical values such as percentage of semantic segmentation undergone by objects in the image, the maximum fine recognition percentage with which an item was detected in the image, and the percentage similarity of an image to the image with the highest number of likes. Using libraries for data handling such as scikit-learn, NumPy, Pandas, data exploration and cleaning was done. These values were entered into the database. The OpenCV model was used and it returned 5 objects in the image with the highest percentage certainty of their identification, their label. Out of the objects, the percentage of the object with the highest recognition certainty was selected. Data cleansing offers a better data quality which will be a great help for the organization to make sure their data is ready for the analyzing phase. Images with human faces were fed to Google Vision API, which returned the percentage certainty with which the model predicted the labels of various objects, textures, colors it identified in the picture.

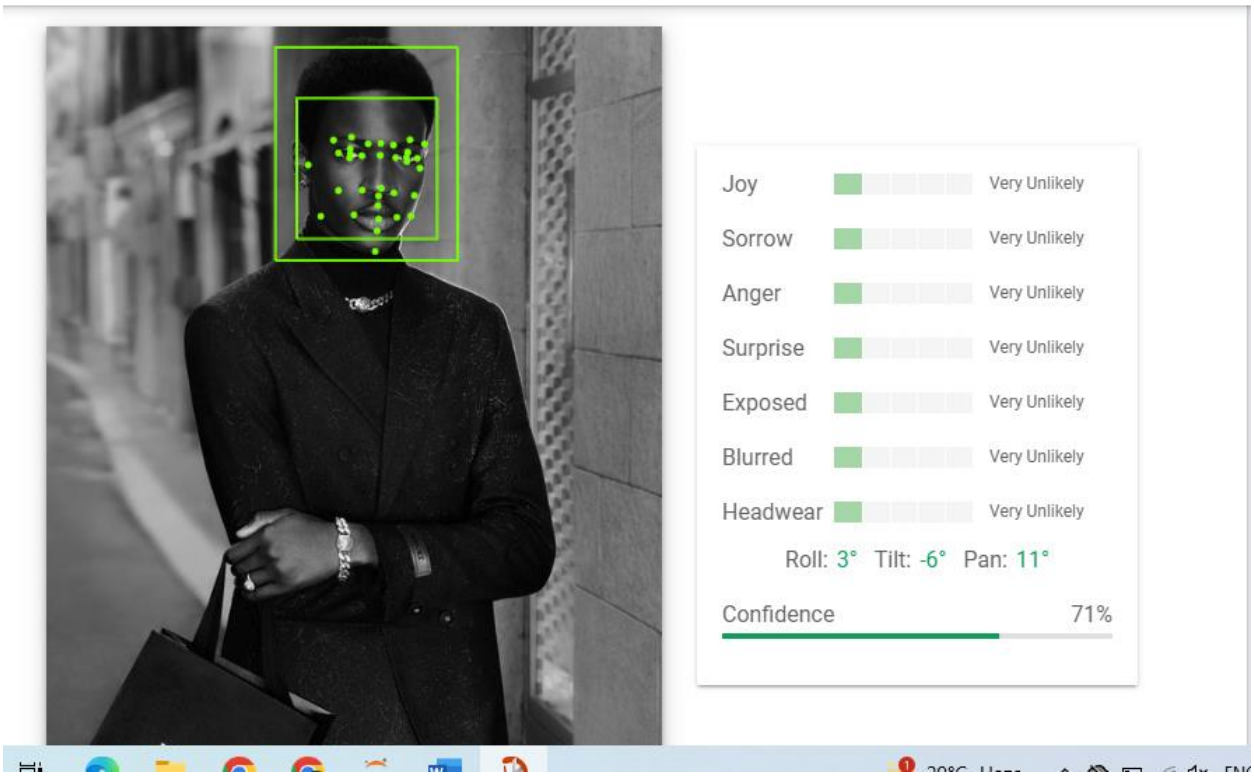

Fig 1.1: Image from dataset on which Google Vision API applies object identification

Emotion	Possibility
Joy	Very Unlikely
Sorrow	Very Unlikely
Anger	Very Unlikely
Surprised	Very Unlikely
Exposed	Very Unlikely
Blurred	Very Unlikely
Headwear	Very Unlikely
Roll	3°
Tilt	-6°
Pan	11°

Table 2.1: Output obtained by applying object identification on an image form sample dataset using Google Vision API

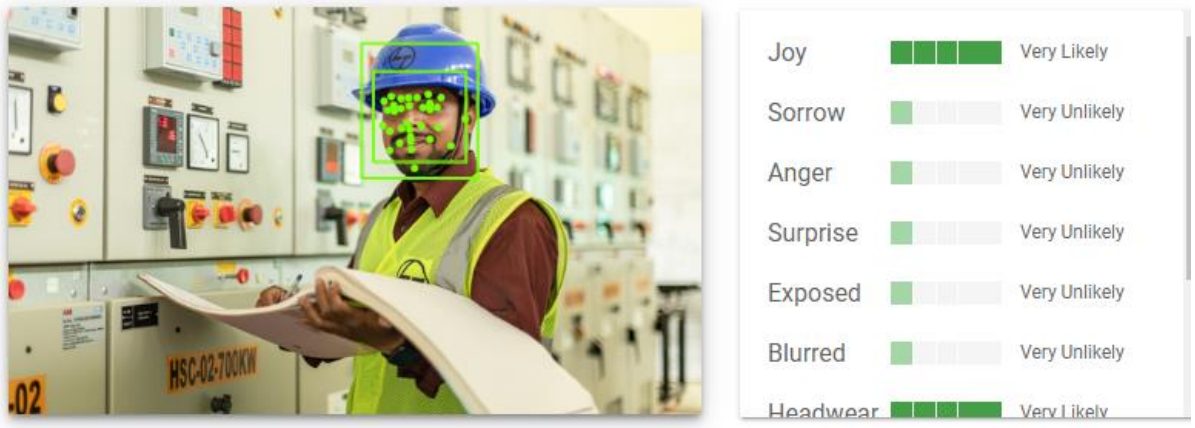

Fig 1.2

Emotion	Likeliness
Joy	Very Likely
Sorrow	Very Unlikely
Anger	Very Unlikely
Surprise	Very Unlikely
Exposed	Very Unlikely
Blurred	Very Unlikely
Headwear	Very Likely
Roll	-3°
Tilt	7°
Pan	-16°
Confidence	98%

Table 2.2: Output obtained by applying object identification on an image form sample dataset using Google Vision API

The Roll, Tilt, and Pan values provided by Google Vision are part of the face detection and analysis feature. These values describe the orientation of a detected face in the image, which can provide context for interpreting the emotional expression.
Here's a detailed interpretation of each of these values:

1. **Roll (3°)**: Roll refers to the rotation of the face around the line of sight. A roll of 3° indicates that the face is slightly rotated clockwise (to the right) when looking directly at it. This small angle suggests that the face is almost upright, with minimal tilt to either side.
2. **Tilt (-6°)**: Tilt refers to the up or down angle of the face. A tilt of -6° means that the face is tilted slightly downward. This downward tilt can sometimes be associated with emotions like sadness, shame, or thoughtfulness, but it is relatively slight, so the impact on emotional interpretation might be minimal.
3. **Pan (11°)**: Pan refers to the side-to-side rotation of the face. A pan of 11° indicates that the face is turned slightly to the right. This slight turning of the face can influence the visibility of facial features and can suggest the person is looking slightly to the side rather than straight ahead.

Implications for Emotion Identification:
The orientation of the face can affect how certain emotions are perceived. Here's how these specific angles might influence emotion identification:

- **Neutral**: The angles provided are relatively minor, suggesting that the face is mostly oriented towards the camera. Minor deviations like these often still allow for clear visibility of facial expressions, making it easier to identify emotions accurately.
- **Positive Emotions (e.g., Happiness, Surprise)**: These emotions often involve wide, symmetrical features such as a smile or wide eyes. The slight pan to the right and the small roll should not significantly obscure these features. However, the slight downward tilt might make the lower part of the face more prominent.
- **Negative Emotions (e.g., Sadness, Anger)**: Emotions like sadness might be slightly enhanced by the downward tilt, as a lowered head can be associated with sadness or introspection. Anger might still be identifiable through features like furrowed brows, even with the slight head tilt and turn.
- **Subtle Emotions (e.g., Contempt, Fear)**: These emotions might be harder to detect if the orientation obscures subtle facial cues. For instance, a slight turn might obscure one side of the mouth, which could be critical for detecting a sneer (contempt) or a slight downturn (fear).

The face orientation provided (Roll: 3°, Tilt: -6°, Pan: 11°) suggests that the face is almost directly facing the camera with minor deviations. This positioning should generally allow for effective emotion identification, with only slight adjustments needed for the minor head tilt and turn. The small angles imply that most facial features should still be visible and interpretable, allowing for accurate emotion recognition.

Fig 1.3 Image on which Google Vision Api has been applied for object identification

Object	Percentage certainty of identification(%)
Green	92
Lighting	86
Rectangle	84
Magenta	76
Bottle	73
Jewellery	71
Eyewear	65

Table 2.3Object Identification Using Google Vision API

The output given by the Google Vision API identified various items in the image and returned their labels along with the corresponding confidence or level of certainty that the predicted label was correct. A few examples of the identified objects in the table have been included above including colors like green, magenta, lighting, eyewear, jewelry, etc.

4.2.3 Feature selection and Machine learning model application

Machine learning algorithms are used to parallelise the workloads. To achieve paramount performance required parameters are tuned in the algorithms. The jobs are implemented using machine learning techniques using various parameters. The performance is examined by executing various features and verifying time constraints depending on the assignments for a cluster. The attempt is made to obtain minimum execution time using python language for implementing machine learning algorithms. Supervised and unsupervised techniques of machine learning algorithms are used differentiating the performance evaluation and time efficiency. Linear regression, logistic regression and K-means clustering techniques are used to evaluate the data analytic jobs(Lakshmi,2018).

Usually feature selection is done on the basis of confusion matrices and correlation matrices.

5. Results and Interpretation

Machine Learning Model	Accuracy	Mean Squared Error	R squared value
Gradient Boosting Classifier	0.7692	0.2308	0.0714
Random Forest Classifier	0.7179	0.2821	-0.1349
Ensemble Voting Classifier	0.7179	0.2821	-0.1349

Table 2.4 Evaluation metrics for machine learning models

Since Ensemble voting classifiers also use some decision trees and random forests for classification of data for the generation of the required output, their accuracy, mean squared error value and R-squared value area almost identical.

```
Gradient Boosting Classifier:
Accuracy: 0.7692
MSE: 0.2308
R^2: 0.0714
Gradient Boosting Feature Importances:
semantic_seg_perc: 0.1495
max_fine_recog: 0.2599
similarity: 0.5906
Random Forest Classifier:
Accuracy: 0.7179
MSE: 0.2821
R^2: -0.1349
Random Forest Feature Importances:
semantic_seg_perc: 0.2543
max_fine_recog: 0.3077
similarity: 0.4380
Ensemble Voting Classifier:
Accuracy: 0.7179
MSE: 0.2821
R^2: -0.1349
```

Figure 1. Evaluation metrics and feature importances for machine learning classifier models

The above Figure 1 shows the output of the evaluation metric values for each machine learning model that was run on the data. The metric parameters used for evaluation are accuracy level of the predictions made by the models, when compared to the original values of the output, the R-squared value and the mean-squared error value.

Gradient Boosting Classifier

- **Accuracy: 0.7692**
  - This means the model correctly predicts the target variable 76.92% of the time.
- **MSE (Mean Squared Error): 0.2308**
  - This metric measures the average squared difference between the predicted and actual values. Lower MSE indicates better model performance.
- **R^2: 0.0714**
  - R-squared represents the proportion of variance in the dependent variable that is predictable from the independent variables. An R^2 value of 0.0714 indicates that only 7.14% of the variance in the target variable is explained by the model. This suggests that the model does not explain the data well.
- **Feature Importances:**
  - **semantic_seg_perc: 0.1495**
  - **max_fine_recog: 0.2599**
  - **similarity: 0.5906**
  - These values indicate the relative importance of each feature in making predictions. Here, similarity is the most important feature, followed by max_fine_recog and semantic_seg_perc.

Random Forest Classifier

- **Accuracy: 0.7179**
  - The model correctly predicts the target variable 71.79% of the time.
- **MSE (Mean Squared Error): 0.2821**
  - Higher than that of the Gradient Boosting Classifier, indicating it has a higher average squared difference between predicted and actual values.
- **R^2: -0.1349**
  - A negative R^2 value indicates that the model performs worse than a horizontal line (mean of the data). This suggests the model does not fit the data well at all.
- **Feature Importances:**
  - **semantic_seg_perc: 0.2543**
  - **max_fine_recog: 0.3077**
  - **similarity: 0.4380**
  - Here, similarity is still the most important feature but with less dominance compared to Gradient Boosting. max_fine_recog and semantic_seg_perc are also significant.

Ensemble Voting Classifier

- **Accuracy: 0.7179**
  - Same as the Random Forest Classifier, indicating that the voting ensemble's performance is identical to one of its constituent models.
- **MSE (Mean Squared Error): 0.2821**
  - Same as the Random Forest Classifier, suggesting that the ensemble method did not improve the mean squared error.
- **R^2: -0.1349**
  - Same as the Random Forest Classifier, indicating that the ensemble model performs equally poorly in terms of variance explanation.

Summary and Recommendations

1. **Model Performance**:
   - The Gradient Boosting Classifier performs better than the Random Forest and Ensemble Voting Classifiers in terms of accuracy and MSE.
   - The Gradient Boosting Classifier's R^2 value is low but still positive, indicating some predictive power, while the Random Forest and Ensemble models have negative R^2 values, indicating poor performance.
2. **Feature Importance**:
   - In both the Gradient Boosting and Random Forest models, similarity is the most important feature, followed by max_fine_recog and semantic_seg_perc.
   - This suggests that similarity plays a crucial role in the prediction, and more focus should be given to improving and leveraging this feature.
3. **Ensemble Voting Classifier**:
   - The Ensemble Voting Classifier does not provide any improvement over the Random Forest Classifier, indicating that combining the models in this way does not enhance performance.
4. **Further Steps**:
   - Consider tuning the hyperparameters of the Gradient Boosting Classifier to potentially improve its performance further.
   - Investigate additional features or more advanced feature engineering to improve the model's ability to explain variance (increase R^2).

- Evaluate other ensemble methods or stacking approaches to see if combining models differently might yield better results.

Table 2.5 gives the values of mean squared error and R squared value for each machine learning model applied to the given dataset.

Machine Learning Model	Mean Squared Error	R squared value
Linear Regression	4127569.21	-0.56
Polynomial Regression	4019464.31	-0.52
Ridge Regression	2886188.09	-0.09
Lasso Regression	4125781.94	-0.56
Elastic Net	2778119.35	-0.05
Decision Tree	7090142.05	-1.69
Random Forest	4775736.71	-0.81
SVR	2901835.27	-0.10
Poisson Regression	2655751.5051	-0.2886

Table 1.3: Evaluation Metrics for Various Model

Here, the parameters mean squared error and R-squared value have been used as measures of accuracy for the various machine learning models that have been applied to the dataset to predict social media engagement and find how it is related to higher image attributes. For linear regression, for example, the high MSE indicates that the model has a significant error in its predictions. The negative R^2 value suggests that the model performs worse than a simple horizontal line (mean of the target variable), implying poor predictive power. Similar to linear regression, a high MSE indicates substantial prediction errors. The negative R^2, though slightly better than linear regression, still indicates that the model does not capture the underlying relationship well and performs worse than a baseline model. Ridge regression shows a lower MSE compared to linear and polynomial regression, suggesting improved prediction accuracy. The R^2 value, while still negative, is closer to zero, indicating the model performs somewhat better than the others but still lacks sufficient predictive power. Lasso regression's performance is almost identical to linear regression, with a high MSE and a negative R^2, indicating poor model performance.

Elastic Net shows the lowest MSE among the linear models, indicating better predictive accuracy. The R^2 value is the closest to zero, suggesting this model has the best performance among the linear models but still does not explain the variability well. For decision tree model, Very high MSE indicates extremely poor prediction accuracy.

The highly negative R^2 suggests the model performs much worse than a baseline model, indicating overfitting or other issues. Random Forest performs better than the decision tree but still has a high MSE and negative R^2, indicating it does not capture the relationship well in this context. SVR shows a lower MSE than most models, similar to Ridge and Elastic Net.

The R^2 value is slightly negative, indicating marginally better performance than a baseline model but still not adequate.

Overall, all models show high MSE and negative R^2 values, indicating poor performance. Several potential issues could explain these results:

1. **Feature Relevance:** The selected features (% semantic segmentation, fine recognition, similarity) may not be strongly predictive of total social media engagement.

2. **Data Quality:** There might be noise or irrelevant data in the dataset affecting model performance.
3. **Model Complexity:** Some models might be overfitting or underfitting the data.

Recommendations:

•Feature Engineering: Explore additional or alternative features that might better capture the relationship with social media engagement.

•Data Cleaning: Ensure the dataset is clean, free of outliers, and well-prepared.

•Hyperparameter Tuning: Optimize the parameters of models, particularly for complex models like Random Forest and SVR.

•Model Selection: Consider ensemble methods or more advanced models like Gradient Boosting Machines (GBMs) or Neural Networks if simpler models do not perform well.

Given these statistics, further investigation and refinement are needed to build a more accurate predictive model.

The next table shows the equations derived from some of these models.

Machine Learning Model	Equation
Poisson Regression	log(Y) = 5.617310772456509 + (-0.00 * const) + (1.93 * semantic_seg_perc) + (-0.96 * max_fine_recog) + (4.67 * similarity)
Logistic Regression	logit(P(Y=1))=−4.3246+0.5906×semantic_seg_perc−0.3754×max_fine_recog+ 24.4115×similarity
Linear Regression	Y = -6905.7624502963545 + (1373.74 * semantic_seg_perc) + (-357.86 * max_fine_recog) + (43886.96 * similarity)
Ridge Regression	Y = -2395.8245882024157 + (1338.93 * semantic_seg_perc) + (-981.94 * max_fine_recog) + (20233.64 * similarity)
Lasso Regression	Y = -6901.981613450701 + (1371.83 * semantic_seg_perc) + (-356.67 * max_fine_recog) + (43867.66 * similarity)
Elastic Net Regression	Y = 701.879821201625 + (756.07 * semantic_seg_perc) + (-849.10 * max_fine_recog) + (4011.54 * similarity)

Table 2.6: Relation Equations of various models, where Y=number of likes + number of comments + number of shares

The equations given in the above table 1.4 represent different types of regression models (Linear Regression, Ridge Regression, Lasso Regression, and Elastic Net Regression) that predict a dependent variable Y based on three independent variables: semantic_seg_perc, max_fine_recog, and similarity. Each model assigns a different coefficient to these variables, which indicates the strength and direction of their relationship with YYY.

The Poisson Regression equation is:

log⁡(*Y*)=5.6173+(−0.00)·const+1.93·semantic_seg_perc+(−0.96)·max_fine_recog+4.67·similarity log(*Y*)=5.6173+(−0.00)·const+1.93·semantic_seg_perc+(−0.96)·max_fine_recog+4.67·similarity

Where:

- *Y* is the expected value of the dependent variable (total social media engagement).
- semantic_seg_percsemantic_seg_perc is the percentage of semantic segmentation.
- max_fine_recogmax_fine_recog is the maximum fine recognition value.
- similaritysimilarity is the similarity measure.

Coefficients Interpretation

- **Intercept (5.6173):** This is the log of the expected count when all predictors are zero. In terms of the original scale, the expected count is exp⁡(5.6173)≈276.53exp(5.6173)≈276.53.
- **semantic_seg_perc (1.93):** For a one-unit increase in semantic segmentation percentage, the log of the expected count of social media engagement increases by 1.93. This means the expected count increases by a factor of exp⁡(1.93)≈6.89exp(1.93)≈6.89, holding other factors constant.
- **max_fine_recog (-0.96):** For a one-unit increase in maximum fine recognition, the log of the expected count of social media engagement decreases by 0.96. This means the expected count decreases by a factor of exp⁡(−0.96)≈0.38exp(−0.96)≈0.38, holding other factors constant.
- **similarity (4.67):** For a one-unit increase in similarity, the log of the expected count of social media engagement increases by 4.67. This means the expected count increases by a factor of exp⁡(4.67)≈106.8exp(4.67)≈106.8, holding other factors constant.

The equation suggests that higher semantic segmentation and similarity values are associated with higher social media engagement, while higher fine recognition values are associated with lower engagement. The large positive coefficient for similarity indicates a strong positive impact on engagement, while the negative coefficient for fine recognition suggests a negative impact.
The high MSE indicates that there is still a significant error in the model's predictions.
The negative R^2 suggests that the model is not effective in explaining the variability in social media engagement. It performs worse than a simple mean-based model.

Recommendations:

1. **Feature Engineering:** Consider creating or exploring additional features that might better capture the relationship with social media engagement.
2. **Model Validation:** Perform cross-validation to ensure the model's stability and to check for overfitting.
3. **Alternative Models:** Explore other regression techniques, such as Zero-Inflated Poisson, Negative Binomial, or even non-linear models, to capture potential complexities in the data.
4. **Data Examination:** Re-examine the data for outliers, multicollinearity, or other issues that might affect model performance.

For logistic regression, here is the summary of the evaluation metrics:

1.) Accuracy: 0.4615 - The model correctly classifies 46.15% of the cases. This indicates a poor performance as it is close to random guessing (50%).
2.) MSE: 0.5385 - The Mean Squared Error indicates the average squared difference between the observed actual outcomes and the outcomes predicted by the model.
3.) R^2: -1.1667 - The negative R^2 value suggests that the model is performing worse than a simple mean model. This means that the logistic regression model does not explain the variance in the engagement data well.

The logistic regression model provides a linear equation on the log-odds scale.
Coefficients and Significance

1.) const (-4.3246): The intercept term. This indicates the log-odds of the baseline when all predictors are zero.

2.) semantic_seg_perc (0.5906): Positive but not statistically significant (p-value = 0.400). This means an increase in semantic segmentation percentage slightly increases the log-odds of higher engagement, but this effect is not statistically significant.
3.) max_fine_recog (-0.3754): Negative and not statistically significant (p-value = 0.565). This suggests that fine recognition may reduce the log-odds of higher engagement, though this effect is not statistically significant.
4.) similarity (24.4115): Positive and highly significant (p-value < 0.001). This indicates that an increase in similarity has a strong positive effect on the log-odds of higher engagement.

The linear regression model's equation gives us the following statistics:

- **Intercept (−6905.76)**: When all independent variables are zero, the predicted Y value is −6905.76.
- **semantic_seg_perc (1373.74)**: For each 1% increase in semantic_seg_perc, Y increases by 1373.741, holding other variables constant.
- **max_fine_recog (−357.86)**: For each unit increase in max_fine_recog, Y decreases by 357.86, holding other variables constant.
- **similarity (43886.96)**: For each unit increase in similarity, Y increases by 43886., holding other variables constant.

For the ridge regression equation

- **Intercept (−2395.82)**: When all independent variables are zero, the predicted Y value is −2395.82.
- **semantic_seg_perc (1338.93)**: For each 1% increase in semantic_seg_perc, Y increases by 1338.93, holding other variables constant.
- **max_fine_recog (−981.94):** For each unit increase in max_fine_recog, Y decreases by 981.94, holding other variables constant.
- **similarity (20233.64)**: For each unit increase in similarity, Y increases by 20233.64, holding other variables constant.

Lasso Regression

- **Intercept (−6901.98)**: When all independent variables are zero, the predicted Y value is −6901.98.
- **semantic_seg_perc (1371.83)**: For each 1% increase in semantic_seg_perc, Y increases by 1371.83, holding other variables constant.
- **max_fine_recog (−356.67**For each unit increase in max_fine_recog, Y decreases by 356.67, holding other variables constant.
- **similarity (43867.66)**: For each unit increase in similarity, Y increases by 43867.66, holding other variables constant.

Elastic Net Regression

- **Intercept (701.88)**: When all independent variables are zero, the predicted Y value is 701.88.
- **semantic_seg_perc (756.07)**: For each 1% increase in semantic_seg_perc, Y increases by 756.07, holding other variables constant.
- **max_fine_recog (−849.10)**: For each unit increase in max_fine_recog, Y decreases by 849.10849.10849.10, holding other variables constant.
- **similarity (4011.54)**: For each unit increase in similarity, Y increases by 4011.54, holding other variables constant.

Overall Comparison

1. **Semantic Segmentation Percentage (semantic_seg_perc)**:
    - The positive coefficients in all models indicate that an increase in semantic_seg_perc leads to an increase in Y, though the magnitude of this effect varies slightly across models.
2. **Maximum Fine Recognition (max_fine_recog)**:
    - The negative coefficients in all models suggest that an increase in max_fine_recog decreases Y. The effect is strongest in the Ridge and Elastic Net models.
3. **Similarity (similarity)**:
    - The positive coefficients indicate a strong positive relationship between similarity and Y. The Linear and Lasso Regression models show a very high impact of similarity on Y, while the Ridge and Elastic Net models show a less pronounced but still significant effect.

These differences arise due to the different regularization techniques used in each regression method:

- **Linear Regression**: No regularization, purely minimizes the error.
- **Ridge Regression**: Adds a penalty for larger coefficients (L2 regularization), which tends to shrink coefficients.
- **Lasso Regression**: Adds a penalty that can shrink coefficients to zero (L1 regularization), effectively performing feature selection.
- **Elastic Net Regression**: Combines both L1 and L2 regularization, balancing between shrinking coefficients and feature selection.

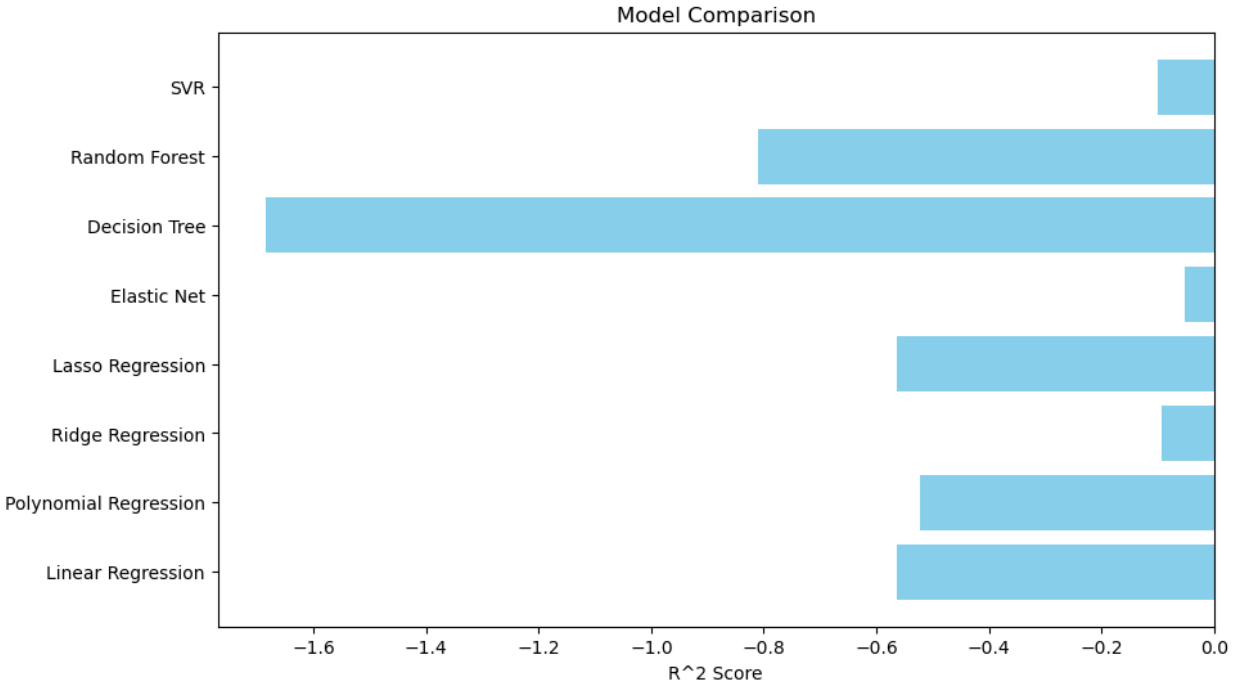

Table1. Graphical comparison of R-Squared values of various models

The above table shows the values of R square for various machine learning models used to compare the data, and classify it in order to generate the output. A bar is shown comparing the R^2

(R-squared) scores of various machine learning models. The R^2 score is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Here's a detailed interpretation:

1. **Model Performance**: The graph shows the performance of different models in terms of their R^2 scores.
2. **Negative R^2 Scores**: All the models have negative R^2 scores, which indicates that these models are performing worse than a horizontal line (mean of the target values). In other words, these models are not fitting the data well at all.
3. **Model Comparisons**:
    - **Decision Tree**: This model has the worst performance with the lowest R^2 score of around -1.6.
    - **Random Forest**: This model also performs poorly with an R^2 score of around -1.1.
    - **SVR (Support Vector Regression)**: This model has the least negative R^2 score, around -0.1, indicating it is the least poorly performing model among the compared models.
    - **Elastic Net**: This model has an R^2 score slightly below zero.
    - **Lasso Regression**: This model has a negative R^2 score of approximately -0.5.
    - **Ridge Regression**: This model has a negative R^2 score of around -0.2.
    - **Polynomial Regression**: This model has an R^2 score of approximately -0.4.
    - **Linear Regression**: This model has an R^2 score of around -0.3.
4. **Implications**:
    - **Decision Tree and Random Forest**: These models might be overfitting or underfitting the data, resulting in very poor performance.
    - **SVR**: This model, despite still having a negative R^2 score, is the best among the compared models, suggesting it might be more suitable for this particular dataset, though still not ideal.
    - **Regression Models**: The regression models (Elastic Net, Lasso, Ridge, Polynomial, and Linear Regression) all show poor performance, indicating that the linear assumptions might not be suitable for this dataset.
5. **Actionable Insights**:
    - Consider revisiting the data preprocessing steps to ensure data quality.
    - Experiment with other types of models or tune the hyperparameters of the existing models.
    - Investigate the dataset for any anomalies or patterns that might be affecting the model performance.
    - Try ensemble methods, boosting, or different architectures to improve the performance.

In summary, all models are underperforming for this particular dataset as indicated by their negative R^2 scores. The SVR model, while still not ideal, shows relatively better performance compared to the others. Further model tuning or trying different algorithms may be necessary to improve prediction accuracy.

Logit Regression Results

Dep. Variable:	total_engage	No. Observations:	152
Model:	Logit	Df Residuals:	148
Method:	MLE	Df Model:	3
Date:	Tue, 28 May 2024	Pseudo R-squ.:	0.1081
Time:	22:39:47	Log-Likelihood:	-93.954
converged:	True	LL-Null:	-105.35
Covariance Type:	nonrobust	LLR p-value:	4.481e-05

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.3246	1.165	-3.714	0.000	-6.607	-2.042
semantic_seg_perc	0.5906	0.702	0.841	0.400	-0.786	1.967
max_fine_recog	-0.3754	0.653	-0.575	0.565	-1.656	0.905
similarity	24.4115	6.115	3.992	0.000	12.427	36.396

Table 1. Statistics and Evaluation metrics of the Logistic Regression Model

1. The field names represent the following quantities:
   coef: These are the estimated coefficients for the regression equation. They represent the average change in the dependent variable for a one-unit change in the predictor, holding other predictors constant.
2. **std err**: These are the standard errors of the coefficients, indicating the average amount by which the estimated coefficient would vary if the model were re-estimated with a different sample.
3. **z**: These are the z-values (coefficients divided by their standard errors). They indicate how many standard deviations the coefficient is away from zero.
4. **P>|z|**: These are the p-values for the z-tests. They indicate the probability of observing a z-value as extreme as the one observed, assuming the null hypothesis (that the coefficient is zero) is true.
5. **[0.025, 0.975]**: These are the 95% confidence intervals for the coefficients. They provide a range of values within which the true coefficient is likely to fall with 95% confidence.

Detailed Interpretation of Each Coefficient

1.) **Const**: The intercept (constant) is -4.3246, which is the predicted value of the dependent variable when all predictors are zero. The p-value is very low (<0.001), indicating that this coefficient is significantly different from zero. The 95% confidence interval does not include zero, reinforcing this significance.

2.) **Semantic_seg_perc**: The coefficient for semantic_seg_perc is 0.5906, suggesting a positive relationship with the dependent variable. However, the p-value (0.400) indicates that this relationship is not statistically significant at the 0.05 level. The 95% confidence interval includes zero, further indicating that this predictor might not have a significant effect.

3.) **Max_fine_recog**: The coefficient for max_fine_recog is -0.3754, suggesting a negative relationship with the dependent variable. However, the p-value (0.565) indicates that this relationship is not statistically significant at the 0.05 level. The 95% confidence interval includes zero, further suggesting that this predictor might not have a significant effect.

4.) **Similarity**: The coefficient for similarity is 24.4115, suggesting a strong positive relationship with the dependent variable. The p-value is very low (<0.001), indicating

that this relationship is statistically significant. The 95% confidence interval does not include zero, reinforcing the significance of this predictor.

The intercept and similarity coefficients are statistically significant, with very low p-values, indicating strong evidence that these predictors affect the dependent variable.

The semantic_seg_perc and max_fine_recog coefficients are not statistically significant, as indicated by their high p-values and confidence intervals that include zero. This suggests that these predictors might not have a meaningful impact on the dependent variable in this model.

6.Conclusion and Discussion

As evidenced by the above data, while all the models and the dataset require further fine tuning by methods such as hyperparameter tuning, changing the sample size and so on. One fact does come to light as evidenced by the research experiment conducted. The amount of social media engagement received by a post containing an image depends on the image's similarity to another image included in a post has very high engagement. So, the similarity to the image with the highest number of likes is the factor that influences the social media engagement a post receives followed by the maximum percentage of semantic segmentation and the maximum percentage of fine recognition. This can be evidenced as seen from the evaluation of the created database of images by various machine learning models such as elastic regression, lasso regression, etc. Here the independent variables would be the higher image attribute values calculated by image mining techniques such as YOLO, vgg_face model, etc. to name a few. The dependent variable is the social media engagement which is the number of likes, comments and shares. This is a count variable and hence can be estimated using the models like Poisson regression and negative binomial regression. While none of the models are entirely accurate, however the support vector (machine) regression model, and the gradient boosting model seems to perform remarkably better in comparison to the other models. The linear regression model and Poisson model equations both reveal that the similarity to an image post with the highest social media engagement

7.Limitations

Due to the small sample size of data chosen for evaluation, many of the models do not give accurate performance. This could be due to inconsistencies in the data, or not choosing a large enough sample size of the database to train the model

8.Implications and Future Direction

6.Conclusion

References:

1. Ravi, Dr & Sujaya, Mr & Scholar, Kumar & Sabri, Mustafa. (2021). SOCIAL MEDIA MARKETING: A CONCEPTUAL STUDY. SSRN Electronic Journal. 8. 63-71.
2. Brodie, R. J., Ilic, A., Juric, B., & Hollebeek, L. (2013). Consumer engagement in a virtual brand community: An exploratory analysis. Journal of Business Research, 66(1), 105–114. doi:10.1016/j. jbusres.2011.07.029
3. Sashi, C. M. (2012). Customer engagement, buyer-seller relationships, and social media. Management Decision, 50(2), 253–272. doi:10.1108/02651330810851872

4. Chen, X., and Qasim, H. (2021). Does E-Brand experience matter in the consumer market? Explaining the impact of social media marketing activities on consumer-based brand equity and love. J. Consumer Behav. 20, 1065–1077. doi: 10.1002/cb.1915
5. Jamil Khalid, Dunnan Liu , Gul Rana Faizan , Shehzad Muhammad Usman , Gillani Syed Hussain Mustafa , Awan Fazal Hussain,(2022) Role of Social Media Marketing Activities in Influencing Customer Intentions: A Perspective of a New Emerging Era, Frontiers in Psychology, Volume-12 DOI=10.3389/fpsyg.2021.808525
6. Ebrahim, R. S. (2020). The role of trust in understanding the impact of social media marketing on brand equity and brand loyalty. J. Relat. Marke. 19, 287–308. doi: 10.1080/15332667.2019.1705742
7. Chi, Hsu-Hsien. 2011. "Interactive Digital Advertising VS. Virtual Brand Community: Exploratory Study of User Motivation and Social Media Marketing Responses in Taiwan." Journal of Interactive Advertising 12: 44-61.
8. Mangold, Glynn W., and David J. Faulds. 2009. "Social Media: The New Hybrid Element of the Promotion Mix." Business Horizons 52: 357-365.
9. Nadaraja, Rubathee & Yazdanifard, Assoc. Prof. Dr. Rashad. (2013). Social Media Marketing SOCIAL MEDIA MARKETING: ADVANTAGES AND DISADVANTAGES.
10. Chen, Yan & Sherren, Kate & Smit, Michael & Lee, Kyung. (2021). Using social media images as data in social science research. New Media & Society. 25. 146144482110387. 10.1177/14614448211038761
11. Ravi, Dr & Sujaya, Mr & Scholar, Kumar & Sabri, Mustafa. (2021). SOCIAL MEDIA MARKETING: A CONCEPTUAL STUDY. SSRN Electronic Journal. 8. 63-71.
12. Sashi, C. M. (2012). Customer engagement, buyer-seller relationships, and social media. Management Decision, 50(2), 253–272. doi:10.1108/02651330810851872
13. Chen, X., and Qasim, H. (2021). Does E-Brand experience matter in the consumer market? Explaining the impact of social media marketing activities on consumer-based brand equity and love. J. Consumer Behav. 20, 1065–1077. doi: 10.1002/cb.1915
14. Jamil Khalid, Dunnan Liu , Gul Rana Faizan , Shehzad Muhammad Usman , Gillani Syed Hussain Mustafa , Awan Fazal Hussain,(2022) Role of Social Media Marketing Activities in Influencing Customer Intentions: A Perspective of a New Emerging Era, Frontiers in Psychology, Volume-12 DOI=10.3389/fpsyg.2021.808525
15. Ebrahim, R. S. (2020). The role of trust in understanding the impact of social media marketing on brand equity and brand loyalty. J. Relat. Marke. 19, 287–308. doi: 10.1080/15332667.2019.1705742
16. Chi, Hsu-Hsien. 2011. "Interactive Digital Advertising VS. Virtual Brand Community: Exploratory Study of User Motivation and Social Media Marketing Responses in Taiwan." Journal of Interactive Advertising 12: 44-61.
17. Mangold, Glynn W., and David J. Faulds. 2009. "Social Media: The New Hybrid Element of the Promotion Mix." Business Horizons 52: 357-365.
18. Nadaraja, Rubathee & Yazdanifard, Assoc. Prof. Dr. Rashad. (2013). Social Media Marketing SOCIAL MEDIA MARKETING: ADVANTAGES AND DISADVANTAGES.
19. K. E. (1995). What Theory is Not, Theorizing Is. *Administrative Science Quarterly*, *40*(3), 385–390. https://doi.org/10.2307/2393789

20. Savakis, Andreas & Etz, Stephen & Loui, Alexander. (2000). Evaluation of image appeal in consumer photography. Human Vision and Electronic Imaging V. 3959. 10.1117/12.387147.
21. Singh, S., Gandhi, M., Kar, A.K. and Tikkiwal, V.A. (2023), "How should B2B firms create image content for high social media engagement? A multimodal analysis", *Industrial Management & Data Systems*, Vol. 123 No. 7, pp. 1961-1981. https://doi.org/10.1108/IMDS-08-2022-0470
22. Wukich, C. (2022), "Social media engagement forms in government: a structure-content framework", Government Information Quarterly, Vol. 39 No. 2, 101684.
1. Simon, F. and Tossan, V. (2018), "Does brand-consumer social sharing matter? A relational framework of customer engagement to brand-hosted social media", Journal of Business Research, Vol. 85, pp. 175-184.
2. Rutter, R.N., Barnes, S.J., Roper, S., Nadeau, J. and Lettice, F. (2021), "Social media influencers, product placement and network engagement: using AI image analysis to empirically test relationships", Industrial Management and Data Systems, Vol. 121 No. 12, pp. 2387-2410.
3. Benton, L. (2017), "Five reasons why B2B companies fail at content and social media marketing", B2B News Networks, available at: https://www.b2bnn.com/2017/09/five-reasons-b2b-companies-failcontent-social-media-marketing/ (accessed 15 May 2022).
4. Shao, K. and Janssens, M. (2022), "Who is the Responsible Corporation? A multimodal analysis of power in CSR videos of multinational companies", Organization Studies, Vol. 43 No. 8, pp. 1197-1221.
5. Han, R., Lam, H.K., Zhan, Y., Wang, Y., Dwivedi, Y.K. and Tan, K.H. (2021), "Artificial intelligence in business-to-business marketing: a bibliometric analysis of current research status, development and future directions", Industrial Management and Data Systems, Vol. 121 No. 12, pp. 2467-2497.
6. Caballero, J., Gomez, G., Matic, S., Sanchez, G., Sebastian, S. and Villacanas, A. (2023), ~ "The rise of GoodFATR: a novel accuracy comparison methodology for indicator extraction tools", Future Generation Computer Systems, Vol. 144, pp. 74-89.
7. Arpan Kumar Kar, Yogesh K. Dwivedi, Theory building with big data-driven research - Moving away from the 'What' towards the 'Why', International Journal of Information Management, Volume 54, 2020, 102205, ISSN 0268-4012,
8. Ravi, Dr & Sujaya, Mr & Scholar, Kumar & Sabri, Mustafa. (2021). SOCIAL MEDIA MARKETING: A CONCEPTUAL STUDY. SSRN Electronic Journal. 8. 63-71.
9. Weick, K. E. (1995). What Theory is Not, Theorizing Is. *Administrative Science Quarterly*, *40*(3), 385–390. https://doi.org/10.2307/2393789
10. Savakis, Andreas & Etz, Stephen & Loui, Alexander. (2000). Evaluation of image appeal in consumer photography. Human Vision and Electronic Imaging V. 3959. 10.1117/12.387147.
11. Singh, S., Gandhi, M., Kar, A.K. and Tikkiwal, V.A. (2023), "How should B2B firms create image content for high social media engagement? A multimodal analysis", *Industrial Management & Data Systems*, Vol. 123 No. 7, pp. 1961-1981. https://doi.org/10.1108/IMDS-08-2022-0470
12. Trunfio, M., Rossi, S. Conceptualising and measuring social media engagement: A systematic literature review. *Ital. J. Mark.* **2021**, 267–292 (2021). https://doi.org/10.1007/s43039-021-00035-8

13. Bowden, J. (2009). The process of customer engagement: A conceptual framework. *Journal of Marketing Theory and Practice, 17*(1), 63–74.
14. Brodie, R. J., Hollebeek, L. D., Jurić, B., & Ilić, A. (2011). Customer engagement: Conceptual domain, fundamental propositions, and implications for research. *Journal of Service Research, 14*(3), 252–271.
15. Vivek, S. D., Beatty, S. E., & Morgan, R. M. (2012). Customer engagement: Exploring customer relationships beyond purchase. *Journal of Marketing Theory and Practice, 20*(2), 122–146.
16. Kumar, V., Rajan, B., Gupta, S., & Dalla Pozza, I. (2019). Customer engagement in service. *Journal of the Academy of Marketing Science, 47*(1), 138–160. https://doi.org/10.1007/s11747-017-0565-2
17. Rathnayake, C., & Ntalla, I. (2020). "Visual Affluence" in Social Photography: Applicability of Image Segmentation as a Visually Oriented Approach to Study Instagram Hashtags. *Social Media + Society*, *6*(2). https://doi.org/10.1177/2056305120924758
18. Rosenholtz R., Li Y., Nakano L. (2007). Measuring visual clutter. *Journal of Vision*, 7(2), 1–22.
19. Li, Y., & Xie, Y. (2020). Is a Picture Worth a Thousand Words? An Empirical Study of Image Content and Social Media Engagement. *Journal of Marketing Research*, *57*(1), 1-19. https://doi.org/10.1177/0022243719881113
20. Hagtvedt Henrik, Patrick Vanessa M. (2008), "Air Infusion, The Influence of Visual Art on the Perception and Evaluation of Consumer Products," *Journal of Marketing Research*, 45 (3), 379–89.
21. Zhang Shunyuan, Lee Dokyun, Singh Param Vir, Srinivasan Kannan (2017), "How Much Is an Image Worth? Airbnb Property Demand Estimation Leveraging Large Scale Image Analytics," working paper, https://ssrn.com/abstract=2976021.
22. J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shapeand context modeling for multi-class object recognition and segmentation. In ECCV, 2006.
23. Masahiro Ono, Brandon Rothrock, Yumi Iwashita, Shoya Higa, Virisha Timmaraju, Sami Sahnoune, Dicong Qiu, Tanvir Islam, Annie Didier, Christopher Laporte, Deegan Atha, Vivian Sun, brFlynn Chen, Bhavin Shah, Kathryn Stack, Chris Mattmann, Chapter 9 - Machine learning for planetary rovers☆☆Government sponsorship acknowledged. Contribution prepared by the Contributor on behalf of JPL/Caltech., Machine Learning for Planetary Science, Elsevier, 2022, Pages 169-191, ISBN 9780128187210, https://doi.org/10.1016/B978-0-12-818721-0.00019-7.
24. Tan, Wai & Lim, Tongming. (2020). A Critical Review on Engagement Rate and Pattern on Social Media Sites. 58-61. 10.56453/icdxa.2020.1002.
25. Kumar, V., Rajan, B., Gupta, S., & Dalla Pozza, I. (2019). Customer engagement in service. Journal of the Academy of Marketing Science, 47(1), 138–160. https://doi.org/10.1007/s11747-017-0565-2
26. Keco, D., Obucic, E., & Poturak, M. (2024). Improving the prediction of social media engagement in universities by utilizing feature selection in machine learning. *International Journal of Research in Business and Social Science (2147-4478)*, *13*(1), 372–380. https://doi.org/10.20525/ijrbs.v13i1.3132
27. Gkikas, D.C., Theodoridis, P.K. (2024). How Data Mining is Used in Social Media. Key Performance Indicators' Impact on Image Post Data Characteristics for Maximum User Engagement. In: Kavoura, A., Borges-Tiago, T., Tiago, F. (eds) Strategic Innovative Marketing and Tourism. ICSIMAT 2023. Springer Proceedings in Business and Economics. Springer, Cham. https://doi.org/10.1007/978-3-031-51038-0_50

28. Sinh My, N., Nguyen, L.T.V. and Pham, H.C. (2024), "An integrated model of social media brand engagement: an empirical study of the Vietnamese luxury residential property market", *Asia Pacific Journal of Marketing and Logistics*, Vol. 36 No. 5, pp. 1270-1295. https://doi.org/10.1108/APJML-01-2023-0061
29. Witten, I. H., & Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. SIGMOD Rec., 31(1), 76–77.
30. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
31. Murphy, K. P. (2018). Machine learning: A probabilistic perspective (adaptive computation and machine learning series). The MIT Press: London, UK. https://www.academia.edu/download/62984186/Machine-Learning-A-Probabilistic-PerspectiveAdaptive-Computation-And-Machine-Learning-Series-by20200416-47298-618w08.pdf
32. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer Science & Business Media.
33. Ridzuan, Fakhitah & Zainon, Wan Mohd Nazmee. (2019). A Review on Data Cleansing Methods for Big Data. Procedia Computer Science. 161. 731-738. 10.1016/j.procs.2019.11.177.
34. Wang, Hongzhi, Mingda Li, Yingyi Bu, Jianzhong Li, Hong Gao, and Jiacheng Zhang. (2014) "Cleanix: A Big Data Cleaning Parfait", in the Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, Shanghai, China
35. Crowdflower. (2016) "Data Science Report." Crowdflower. Available from: https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf. [Accessed Jan., 28, 2018].
36. Liu, L., Wiliem, A., Chen, S., & Lovell, B. C. (2017). What is the best way for extracting meaningful attributes from pictures?. *Pattern Recognition*, *64*, 314-326.
37. Côté, P. O., Nikanjam, A., Ahmed, N., Humeniuk, D., & Khomh, F. (2024). Data cleaning and machine learning: a systematic literature review. *Automated Software Engineering*, *31*(2), 54.
38. Jupudi, Lakshmi. (2018). Machine learning techniques using python for data analysis in performance evaluation. International Journal of Intelligent Systems Technologies and Applications. 17. 3. 10.1504/IJISTA.2018.10012853.
39. Lakshmi, J.V.N. (2016) 'Stochastic gradient descent using linear regression with python', IJA-ERA, Vol. 2, No. 8, December, pp.519–524.
40. Manar, A. and Stephane, P. (2015) 'Machine learning with Python', SIMUREX, October.
41.